

Proteomic diagnostics tested

Concerns about a cancer diagnostics test based on proteomics highlight the threat to public trust in healthcare products where the relevant data are not publicly available — and what can be achieved when they are.

It is easy for critics to slam other scientists who go out on a limb and make big claims for their work. This is the situation facing Emanuel Petricoin and Lance Liotta, two researchers with the US government. The pair led research to develop a ground-breaking diagnostic test for ovarian cancer. Their critics allege that the test is flawed and that corporate entities are trying to rush it to market anyway. The jury is still out on whether the test will prove to be useful in the clinic. But perhaps the most important aspect of the debate is that it would never have arisen if Liotta and Petricoin had not posted their data on the Internet. The episode underscores the crucial importance of readily available public data for scientific progress and, ultimately, for public health.

Liotta and Petricoin believe that their test diagnoses ovarian cancer before the disease progresses to an incurable stage. Their test uses proteomics and involves examining all the proteins in a drop of blood, scanning for a pattern that marks out cancer patients. In a widely hailed study (E. F. Petricoin *et al.* *Lancet* **359**, 572–577; 2002), they claimed that their proteomic analysis was highly effective. They posted the data on which they based their conclusions on a government website, followed by two more data sets, allowing others to re-analyse their work. Other researchers have done just that, and claim to find technical problems so troubling that they question the conclusions of the original *Lancet* paper — and even the validity of any diagnostic test based on proteomics (see page 496).

Petricoin and Liotta have defended their methods and the integrity of proteomics as a diagnostic technique for cancer. They also say that a large clinical trial should be performed before the test is marketed to consumers. But the company that worked with the pair on the *Lancet* paper, Correlogic Systems, has already given two larger companies the rights to market a test called OvaCheck in the United States. Correlogic says that it has worked with its licensees, Quest Diagnostics and the Laboratory Corporation of America, to validate the test and ensure that it will work for women around the country in many different testing conditions. But the companies have not released the data on which they base this conclusion.

If OvaCheck works, it could save many women's lives. But the risks of an imperfect diagnostic test are not slight. Women who receive false positives will undergo needless stress and unnecessary surgery, whereas those who receive false negatives may not receive care in time to save their lives. Before OvaCheck hits the market, Correlogic and its licensees need to publish evidence from a large clinical trial that proves that the test works on samples collected by many different doctors around the country and the world. They should give other scientists the opportunity to examine the data and to send in blinded test samples so that they can determine for themselves whether the test gives accurate results. If it really works, the company should have every confidence that independent researchers will support their determination to speed OvaCheck into the clinic. ■

A question of priority

How to sustain the reliability of the patenting system?

The need for scientists to establish priority for their discoveries is fundamental, but how they do it changes with the years. In the seventeenth century, scientists encrypted their results in anagrams embedded in clearly dated letters sent to colleagues. When Galileo discovered the phases of Venus, he created a Latin anagram whose solution translates as: "The mother of love mimics the Moon's shapes". Desperate to discover his rival's secrets, Johannes Kepler struggled to break the code. He thought he had when he derived a sentence referring to a red spot on Jupiter. Wrong issue, wrong planet and a century ahead of the discovery of the real thing.

Science publishing and patenting nowadays provide more transparent systems for establishing priority, but both are under pressure. In particular, patent offices everywhere have been swamped with applications since the revolutions in information technology, biotechnology and materials sciences, and have struggled to clear mounting backlogs. The European Patent Office (EPO) has coped by recruiting more examiners and by the driving productivity of its staff. But have productivity demands gone too far? Are patent examiners being pushed to work so fast, as they claim (see page 493), that they can't deliver quality patents that would withstand scrutiny?

EPO staff are not complainers by nature, so they should be listened to, particularly at a time when external stakeholders are also voicing concerns. Many observers say that the EPO's commercial

orientation overly favours the applicants as customers to be satisfied. Others point to the registration revenues to national patent offices from EPO patents, and say that this factor discourages governments from pressing for quality over quantity.

But what exactly do we mean by patent quality? Any patent must demonstrate novelty, industrial applicability and the involvement of an inventive step. The EPO has a strong reputation in establishing novelty. But it is alleged to be granting some claims for industrial applicability that are inappropriately broad, and granting patents whose inventive step is debatable. Such concerns are themselves hard to test. It would take years to see how many patents are successfully challenged in appeal. Much more useful, and more provocative for the EPO, is the idea that it submit a selection of recently granted patents for external peer review. In the meantime, the concerns are coming from many directions, and confidence in the patenting system is, for the first time, being shaken. Furthermore, patent professionals complain that the quality of incoming applications is now also low.

The EPO's new president, Alain Pompidou, a French biologist, needs to make a clear statement about the office's commitment to quality as soon as he takes over on 1 July. And he must indicate what measures the EPO is going to take to monitor quality. To ignore so much concern would be counter-productive, and would only fuel the smouldering scepticism. Anagrams, anyone? ■

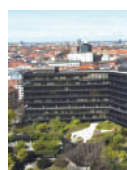




Stem-cell ethics
Korean bioethicists
call for inquiry into
cloning experiment
p490



Lucy's travels
Concern over plans
to exhibit fossil
outside Ethiopia
p491



Quality control
Europe's patent
examiners under
pressure
p493



Virus jump
International
agencies to pool
resources
p494

NASA devolves control of Cassini observations to research teams

Tony Reichhardt, Washington

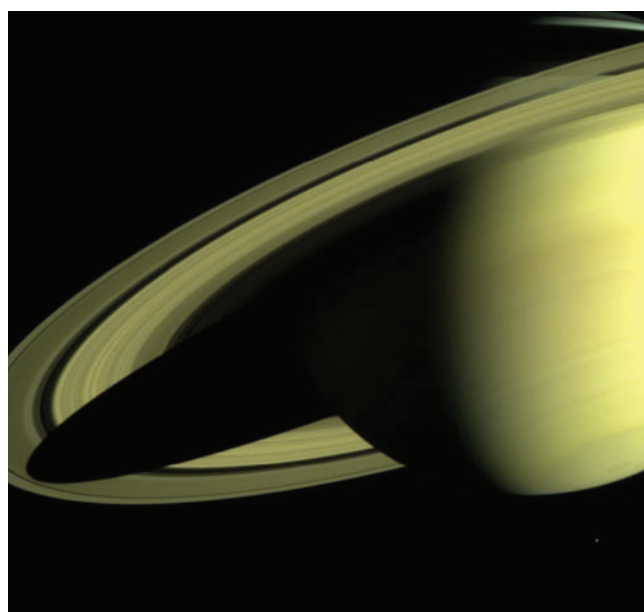
Scientists gearing up for the most ambitious planetary mission in history say it will also be the most demanding for them. For the Cassini mission, NASA has shifted much of the hands-on planning from the agency's own centres to the instrument teams in the universities.

The US\$3.2-billion US–European spacecraft is set to arrive at Saturn on 1 July to begin a four-year, 74-orbit tour of the giant gas planet, its rings and its moons. While a dozen instruments on the main orbiter study the Saturn system close up and map its electromagnetic field, the attached European-built Huygens probe will drop down in January to the surface of Titan, the planet's largest moon.

Similar in design to the Galileo mission that explored Jupiter and its moons in the 1990s, Cassini will be “like Galileo on steroids”, says Torrence Johnson of the Jet Propulsion Laboratory in Pasadena, California, a member of the Cassini imaging team and former Galileo project scientist. Not only will Cassini cram far more observations into half the time, data from the spacecraft will pour in much faster than it did from Galileo, which was hampered by a jammed antenna.

Scientists have had far more responsibility for the details of the Cassini mission than they had with Galileo and with the Voyager missions that surveyed Jupiter, Saturn and the outer planets in the 1970s and 1980s and are now out in deep space. On those projects, investigators who had built cameras and other sensors requested where their instruments should be pointed and when, and technicians at JPL worked out the complex sequences of spacecraft manoeuvres needed to carry out the observations.

But when Cassini was being planned in the early 1990s, NASA decided that instrument teams would work the manoeuvres out for themselves — partly to give the scientists more control and partly to save money. A lot



Getting closer: Saturn as seen from the Cassini spacecraft last month.

of the work was “thrown over the fence” to the outside scientists, says Stamatios Krimigis of the Johns Hopkins University Applied Physics Laboratory in Laurel, Maryland, principal investigator for one of Cassini's magnetospheric experiments.

That allowed JPL to cut its operational budget for Cassini by some 40% compared with that of Galileo, according to Mark Dahl, NASA's programme executive for Cassini. But project scientists complain that they have had to absorb the extra work without adequate funding from NASA.

The problem was compounded by another NASA money-saving decision back in 1992. Instead of mounting the cameras on a separate platform so that they could be directed independently of other instruments, they were fixed to the spacecraft body, making all Cassini's sensors interdependent. This makes it far more difficult to plan the observation manoeuvres. “The complexity of the job was underestimated,” says Krimigis, a sentiment echoed by other Cassini investigators.

Larry Esposito of the University of Colorado in Boulder, who built Cassini's ultraviolet-imaging spectrograph, likes the additional

experimental control that scientists gain by creating their own observing sequences. But, he says, the new strategy carries some risk. While all commands will be checked by JPL before they are sent to Cassini, to ensure that they do not jeopardize the spacecraft, some observations could be less efficient than they might be, as a dozen different teams are involved in the sequencing.

Scientists are reluctant to complain about funding, however, with Cassini's price tag reaching a cool \$3.2 billion. That includes \$400 million for the launch, \$750 million paid by European countries, and \$720 million to operate Cassini and keep the science teams funded until the end of the mission in 2008.

And Cassini's investigators are looking forward to the richest scientific bounty of their careers.

Close study of the intricate ring system should teach researchers about the dynamics of similar disks surrounding other stars. Saturn's moon Enceladus may turn out to be geologically active, with water geysers spouting above its icy surface. Titan, with its constant haze and Earth-like, nitrogen-rich atmosphere, is a mystery that planetary scientists have long wanted to crack. And because Cassini will return hundreds of thousands of images of the planet's swirling clouds, researchers will be able to create extensive movies of the turbulent atmosphere — something they were unable to do at Jupiter because of Galileo's antenna problem.

Cassini entered Saturn's sphere of gravitational influence in mid-May. The action heats up on 11 June, when the inbound spacecraft makes its first close pass of a moon, tiny Phoebe. Then on 1 July, Cassini's main rocket engine will fire for 94 minutes to brake the spacecraft into orbit around Saturn and begin four years of intensive exploration. Comparing the pace of those observations to the United States' best-known car race, Dahl says: “Scientists are looking forward to the start of the Indianapolis 500.”

Double check casts doubt on statistics in published papers

Helen Pearson

A study highlighting statistical gaffes in scientific literature has brought renewed calls for vigilance among mathematically challenged researchers and journal editors.

Statistical tests are sometimes seen as a necessary evil by researchers, who fear their complexity but know that they are needed to test hypotheses. With this aversion in mind, biostatisticians Emili García-Berthou and Carles Alcaraz of the University of Girona, Spain, gauged the extent of statistical errors in four volumes of *Nature* from 2001 (409–412) and a sample of results in two *BMJ* volumes (322–323) from the same year.

The pair used three standard software packages to recalculate 'P values', the parameters by which researchers measure whether a result has statistical significance. Generally a P value of less than 0.05 is taken to be significant and unlikely to have resulted from chance. This may indicate, for example, that blood pressure in a patient group was reduced more by an active drug than by a placebo.

In the *Nature* and *BMJ* papers, each P value was calculated from two other parameters that are included in the papers. García-Berthou and Alcaraz recalculated the P values from these numbers, and found that their results differed from those published in more than 11% of cases. They also found small mistakes, such as rounding errors, in 38% of the *Nature* papers and 25% of the *BMJ* ones (*BMC Med. Res. Methodol.* www.biomedcentral.com/1471-2288/4/13; 2004).

In only 1 case out of 27 did an incorrect P value change a significant result to a non-significant one. But, although minor, some believe that the slip-ups expose a pervasive sloppiness towards statistics in published research. "There are small mistakes that may occasionally have big consequences," says Martin Bland, an expert in medical statistics at the University of York, UK.

Philip Campbell, the editor-in-chief of *Nature*, says the journal will take a closer look at the study's numbers before deciding whether remedial action is needed. He adds that *Nature* has amended its editing practices since the period covered by the study.

Richard Smith, editor of the *BMJ*, says that one way forward is for researchers or journals to publish more raw data on the Internet, where others would be able to check them. ■



Unwatched: questions have been raised over how Woo Suk Hwang's laboratory obtained human eggs.

Korean bioethicists call for inquiry into stem-cell work

David Cyranoski, Tokyo

Bioethicists are pushing for an investigation into the cloning work of a South Korean research team — but are having no luck in finding someone to lead it.

Woo Suk Hwang from Seoul National University and colleagues cloned a human somatic cell, creating an embryo that they used to establish a stem-cell line. It was a huge leap in a field full of medical promise and a major boost for Korean science.

But ethical questions have been raised, particularly about how egg donors were recruited for the experiment and whether they included junior members of the laboratory (see *Nature* 429, 3; 2004).

Hwang denies that any members of his laboratory were donors and he is backed up by the Institutional Review Board (IRB) of Hanyang University Hospital in Seoul, which originally approved the experiment.

But critics question whether the board has been sufficiently rigorous. At its annual meeting in Seoul on 22 May, the Korean Bioethics Association called on Hwang and the review board to answer questions concerning the recruitment of donors and funding sources. "We request the institutes involved and the participants to present clear explanations regarding the following queries about therapeutic human embryonic cloning research," reads the statement they issued.

The questions, posted on the association's homepage and picked up by major South Korean newspapers, include: "Did the Hanyang University Hospital's IRB perform a continuing review on this research project in order to monitor its ethics?"

Some bioethicists accuse the board of not including the non-researchers required by guidelines from the Korea Food and Drug Administration. Institutional review boards are obliged to include "more than one attorney or religious representative, not from the fields of medicine, dentistry, oriental medicine, pharmacy or nursing sciences". Hanyang's consisted of 12 doctors, a pharmacologist, a nurse and a theologian.

The bioethics association wants the case pursued by the National Human Rights Commission, an independent investigative body funded by the government. The commission established a bioethics task force two months ago, but senior commissioner Kyung Seo Park says it was set up to implement rules that come into force next year, not to investigate specific research projects.

"There are no provisions to deal with special cases like Professor Hwang's," Park says. He admits that last month a member of the task force asked the Hanyang review board for documents about the research, and that the board said it was not possible to provide them. But he adds, "The task force does not have any kind of binding power so I told them to stop." Park says the commission may still take up an investigation later.

The situation worries Young Mo Koo, a medical ethicist at the University of Ulsan College of Medicine and member of the bioethics association. The Korea Food and Drug Administration oversees commercial projects, but no one regulates basic research involving human samples, says Koo. "It's a grey area," he adds. "There is a serious need for an investigation." ■

Fossil hunters bristle over plans for US tour

Rex Dalton, San Diego

She may be 3.2 million years old, but plans are being laid to send Lucy on her first vacation. A museum in Texas is negotiating a deal to bring the skeleton of the famous hominid from Ethiopia to the United States for its public debut.

But the idea of transporting the fossil — as opposed to a replicate cast — from east Africa for a three-year travelling exhibition is raising some eyebrows.

Concerned about the exploitation of poorer countries and possible damage to valuable specimens, palaeoanthropologists from 20 nations agreed in 1999 to oppose the transportation of original fossils in travelling exhibitions. The policy, published by the International Association for the Study of Human Paleontology, recommends that “original hominid fossils should not be transported beyond the country of origin, unless there are compelling scientific reasons” for the journey. The International Association of Human Biologists also adopted the policy.

Lucy's original skeleton (*Australopithecus afarensis*) is housed in a vault in the National Museum of Ethiopia in Addis Ababa, where a cast of it is publicly displayed. But under a plan backed by politicians and business leaders in Houston, Texas, and in Ethiopia, the Houston Museum of Natural Science would put the real skeleton on show from 2006.

The exhibition has been conceived to help build business and cultural links between Texas and Ethiopia. On a trip to Addis Ababa in April, a delegation from Texas met with senior government officials, including the president, Girma Wolde-



Dead cert? The original remains of Lucy could be the star attraction in a travelling exhibition.

Giorgis, and obtained a letter of intent from the Ethiopian Tourism Commission that the tour should take place.

But some fossil experts are unhappy with the plan. “I don’t think original fossils should be moved without good scientific reason,” says Bernard Wood, a palaeoanthropologist at George Washington University in Washington DC who signed the 1999 policy. Wood, however, concedes that a lucrative series of US exhibitions could help the

Ethiopian museum. “African museums are badly underfunded,” he says, adding that the exhibition could be justified if enough of the proceeds go back to the museum in Addis Ababa. The terms of the proposed deal have not yet been negotiated.

Alan Parrish, interim president of the Houston museum, says that he wasn’t familiar with the palaeontology association’s policy. “We are aware that there is controversy on the transportation of early artefacts,” he says, adding that the exhibition will be planned with appropriate care, security and reverence, and will be “good for Ethiopia”.

The exhibition would include a range of Ethiopian artefacts and fossils — but the 40%-complete skeleton of Lucy would be its centre-piece. The skeleton was found in 1974 at Hadar by Donald Johanson, now of the Institute of Human Origins at Arizona State University in Tempe. He declined to comment on the exhibition plans.

Anthropologist Mamitu Yilma, director of the Ethiopian museum, says that the agreement is currently “only a letter of understanding”. She adds that “it is not yet decided” if the original Lucy will go on tour. Yilma declined to give a view on it doing so.

But Zeresenay Alemseged, an Ethiopian palaeoanthropologist now at the Max Planck Institute for Evolutionary Anthropology in Leipzig, Germany, said: “I can’t support the exhibit. I want to know more about security and the handling of the fossils.”

Houston museum officials hope to finalize arrangements in the next few months, with a possibility that they may join forces with another major museum in the eastern United States to advance their bid. ■

Scientists deny ethical breach at Kenyan orphanage

Declan Butler

British researchers are hotly contesting allegations that they short-circuited ethical procedures in research involving children at the Nyumbani orphanage in Nairobi, Kenya.

The centre has attracted international interest because some of the children there who have AIDS seem to control their HIV infection without drug therapy, and survive beyond their expected lifespan with a relatively intact immune system.

On 23 May, the Nairobi-based *Sunday Nation* ran a story under the headline “Shame of children used in experiments on Aids”. In it, the paper alleged that Eric Miller, a researcher from the University of Cambridge, carried out research at the orphanage in April without government approval. It also claimed that a group from

the University of Oxford that visited the centre in 2001 had exported blood samples without permission.

Kenya’s National Council for Science and Technology is looking into the allegations. But Ababu Namwamba, a lawyer representing the orphanage, says that the allegations are groundless and originate from a disgruntled former employee. In a statement prepared by Namwamba, the centre says that Miller did no research and only visited the orphanage “on a fact-finding mission” with a view to future research on the impact of food supplements on HIV progression.

The Oxford group’s work, which resulted in two research papers, received ethical clearance from the national council on 8 January 2001, according to the centre. It admits that the original proposal omitted to

request specific consent to send samples to Oxford, but says that once this oversight was spotted, Nyumbani and the university immediately told the council, and got ethical approval in September 2002.

In a letter addressed to the *Nation* last week, the Oxford researchers involved, Sarah Rowland-Jones and Rana Chakraborty from the Medical Research Council’s Human Immunology Unit, added that once they had discovered the discrepancy, they “called an immediate halt to all research and export of samples until the matter could be cleared up”.

The orphanage says that it “experienced the spreading of malicious propaganda”, after it fired a scientist in 2001. It adds that it wants the government to investigate the matter so that it can clear its name. ■

Overseas aid policy needs better science input, inquiry finds

Jim Giles, London

The British government's foreign-aid department looks set to appoint a chief scientist, after being criticized for how it uses science.

Senior academics found fault with the Department for International Development's research programmes at an inquiry begun in January by the House of Commons Science and Technology Select Committee. The department currently spends £80 million (US\$150 million) a year on research in areas such as agriculture and health, with the aim of reducing poverty in poorer nations. Most of the money is distributed as grants to research institutions in Britain and, increasingly, developing countries.

At one of the committee's hearings on 15 March, for example, John Lawton, chief executive of the Natural Environment Research Council, a government funding body, was asked to sum up the department's use of scientific advice in a few words. He chose "complacent", "rather arrogant" and "ill-informed".

Paul Spray, the department's head of research, rejects Lawton's characterization, pointing out that he and other officials meet regularly with people in the research councils to discuss funding programmes. He acknowledges that his staff could interact more with individual scientists, but says a recruitment drive to increase research staff numbers from 7 to 17 should help to rectify this.

The department has asked a three-man task force to examine how it gets advice on science and technology. Sources close to the department and the team, which started work in April and is due to finish in June, say it is likely to recommend the creation of a post of chief scientist. The department will also ramp up its research funding to at least £100 million a year in 2006–07. Its draft research strategy, released on 11 May, identified four priority areas — agricultural productivity in Africa, killer diseases, countries that work in the interests of the poor, and climate change.

The strategy was welcomed by UK overseas development experts, but critics said it included little direct input from scientists — a complaint Spray accepts. He says that in some areas, such as agricultural science, he has been able to identify science organizations that share the department's aim of reducing poverty, but that collaboration in other areas is hampered by a lack of suitable partners. ■



Project member Kelley Lee brandishes one of 40,000 forms submitted to access BAT's depository.

Researchers seize moment to make tobacco data public

Michael Hopkin, London

Public-health researchers have unveiled a project to tackle what they describe as information concealment by the UK-based multinational firm British American Tobacco (BAT). The group aims to publish some 8 million pages of the company's documents on an independent website, making them more easily accessible.

The researchers accuse BAT of obstructing public attempts to access papers at its depository in Guildford, UK, and allege that some files detailing the company's activities have been removed or altered. The facility, they say, limits visitor numbers, doesn't provide an easily searchable index of its material, and does not allow onsite photocopying of documents. Visitors must request copies from BAT, which can take up to 12 months to arrive.

"This sort of conduct raises questions as to the true public availability of the depository's contents," says Kelley Lee, a public-health expert at the London School of Hygiene and Tropical Medicine and a project member.

BAT denies that any files have been deleted. Regarding access to the depository, "it was never designed to work like a public library", says Michael Prideaux, BAT's corporate and regulatory affairs director. He adds that researchers are welcome to reproduce material given to them by BAT.

The depository was opened in 1999 in response to a 1998 settlement between the state of Minnesota and the tobacco industry, which mandated that the documents be made available to the public. In accordance with the ruling, US-based defendants have their Minnesota repository administered by an independent legal firm, but BAT runs its Guildford depository itself.

In 2000, a British government committee recommended that BAT publish the depository's entire contents on the Internet to facilitate access. But so far the company has only stored 350,000 pages electronically, and has not posted those online.

The £2-million (US\$3.6-million) initiative, called the Guildford Archiving Project, is backed by a slew of public-health organizations, including the Wellcome Trust and the American Heart Association. Together they have ordered copies of the depository's entire contents, which involved completing more than 40,000 order forms. They aim to post them, fully indexed, on a website run by the University of California, San Francisco.

The project's leaders claim that, since researchers began visiting the depository in 2000, more than 180 sets of documents, totalling some 36,000 pages, have disappeared from the company's list of files.

In this week's issue of *The Lancet*, a group of the project's researchers describes how documents originally referring to marketing to 16-year-olds were altered to read "18-year-olds" instead (M. E. Muggli, E. M. LeGresley and R. D. Hurt *Lancet* 363, 1812–1819; 2004).

They also claim that an audio recording about a marketing strategy was later deleted, removing phrases such as: "if you just say, this is a cheap cigarette for you dirt poor little black farmers ... they're not going to go for it." The deletion may have been accidental, they add, and a master copy was provided on request. But the incident highlights how some information may go missing, they say.

BAT spokeswoman Teresa La Thangue denies that any documents have been removed, and says that the apparently missing documents were probably duplicates. ■

Pressured staff 'lose faith' in patent quality

Alison Abbott, Munich

Examiners at the European Patent Office (EPO) are losing confidence in its ability to ensure the quality of the patents it issues, according to two separate staff surveys.

In a survey of some 1,300 patent examiners, conducted by the staff union in April, more than three-quarters agreed with the statement that productivity demands from the EPO's managers did not allow them "to enforce the quality standards set by the European Patent Convention". And 90% said that they did not have time to keep up to date with advances in their scientific fields. In a second survey of 730 examiners, done by the EPO itself, only 9% said they believed that the management was "actively involved in improving quality".

But Ciaran McGinley, head of the EPO's controlling office, says that commitment to quality is fundamental to the organization. He adds that the office is currently preparing a thorough report on its approach to quality, which will be presented to staff in the summer.

The number of patent applications processed by the Munich-based office has more than doubled since 1995 to over 160,000 last year. The mean number of hours spent examining each patent claim dropped from 23.8 hours in 1992 to 11.8 hours in 2001, the last year for which figures are available.

But some patent examiners privately contend that the pressure to process files encourages them to approve marginal cases, instead of formulating reasons for rejection. Internal quality controls, in which two other examiners check each patent immediately before it is granted, have also been eroded, they allege. "It is difficult to ask a close colleague to start checking a patent again when you know that he or she has as many files as you have to get through, and just as little time," explains one examiner, who did not want to be identified. In the union survey, two-thirds of respondents said that they



Time out: staff at the European Patent Office say they are too busy to do satisfactory checks.

lacked the time to do back-up examinations "in a satisfactory manner".

Outsiders have questioned the quality of recently granted patents. To qualify for a patent, an invention must demonstrate novelty, industrial applicability and an inventive step—but some are worried about the last, especially in fast-growing areas such as biotechnology.

"The bar for inventive step has been kept too low in genetic patenting," says genome researcher and 2002 Nobel laureate John Sulston of the Sanger Institute in Cambridge, UK. Sulston co-authored a report from the Royal Society published in April 2003, which raised concerns about patent quality. "Officers need plenty of time to consider carefully whether something is really an invention or not," Sulston says.

McGinley says that the EPO has had no complaints. "We've had no feedback from industry about inventive-step issues," he says.

Francis Hagel, an intellectual property manager with the French geoscience-service company CGG, near Paris, argued in April's *Patent World* that there was a "serious quality problem". Hagel also claims that the EPO has followed the commercial orientation of the US Patent and Trademark Office, and is treating its applicants as customers. He says that the EPO should instead be putting more emphasis on determining whether patents deserve to be granted.

Critics point out that the EPO gains fee income from each patent it grants. "Any incentive system that favours patent granting adds pressure to award a patent in marginal cases," says John Pethica, a physicist at Trinity College Dublin and a co-author of the Royal Society report.

McGinley counters, however, that the EPO's approach has cleared a backlog of applications, which he says was "bringing the EPO into disrepute". ■

Brands in peril as Brazil strives to keep AIDS drugs free

Declan Butler

Brazil has pioneered the distribution of free AIDS treatment. But as costs soar, it is considering the issue of compulsory licences to bypass patents on some AIDS drugs.

About 125,000 Brazilians with AIDS take advantage of the nation's 1996 free-drugs policy, and mortality has dropped from 70% to 40%. But the country's generic drugs are becoming obsolete, forcing it to import costly, newer ones.

The Trade-Related Aspects of Intellectual Property Rights (TRIPS) agreement, which

Brazil joined in 1996, prevents it from making generic copies of drugs developed since then. Patient demand for newer therapies means that 8 of the 15 approved drugs are now brand-name imports, and account for 70% of its AIDS treatment budget.

A meeting of Latin American politicians and scientists in Brasilia on 23–24 June will discuss how to assure the long-term viability of free antiretrovirals.

Brazil could invoke a compulsory licence, which would allow it to make generic versions of recent drugs. Under World Trade

Organization rules adopted last year, this is legal in cases of 'national emergency'. But Brazil fears a trade dispute with the United States if it does this, says Michel Lotrowska, from the Rio de Janeiro office of Médecins Sans Frontières. Trade pressure against doing so is very strong, he says.

The threat of such licensing has lowered prices in the past, but it becomes less convincing as Brazil's production of generic drugs falls, and the expertise to make them is lost. So, Brazil will "need to think about compulsory licences", Lotrowska says. ■

Virtual observatory finds black holes in previous data

Munich A team of European astronomers has discovered dozens of supermassive black holes — without using a telescope.

Instead, the team used a ‘virtual observatory’ that searches through images previously taken by the Hubble Space Telescope, the Very Large Telescope in Chile and the Chandra X-ray Observatory. The computer program, called the Astrophysical Virtual Observatory, was able to identify 30 very faint, distant black holes by comparing X-ray and visible light images. The results are set to appear in the journal *Astronomy & Astrophysics*.

Only nine of this particular class of black hole have previously been found, according to lead author Paolo Padovani of the European Southern Observatory in Garching, Germany. The new finds may force theorists to reconsider their previous estimates of how many black holes there are in the Galaxy, he says.

Supporters of virtual observatories say they help to boost the powers of existing telescopes, and give astronomers in parts of the world without access to telescopes the chance to conduct cutting-edge research.



Hen hazard: bird viruses can potentially cross the species barrier to affect humans.

Global agencies join forces against animal diseases

Rome After a spate of diseases that have passed from animals to humans, and with the rise in such diseases spreading between countries, the Food and Agriculture Organization of the United Nations and the World Organisation for Animal Health have banded together in a bid to control future outbreaks.

Many researchers agree that the next major pandemic could result from a virus jumping the species barrier between animals and humans. Such viruses — including HIV, severe acute respiratory syndrome (SARS) and some influenza viruses — are often

deadly because humans have not had the chance to develop resistance to them. The last major flu epidemic, in 1918, which killed more than 40 million people worldwide, was probably caused by a bird virus.

On 24 May, the two organizations agreed to combine their resources and establish a global information network. “Rapid distribution of information and an improved coordination between countries is essential to control contagious diseases,” the two agencies said.

US picks fellows for ‘fresh’ science advice

Washington A group of five scientists poised to advise the US Department of State will “bring fresh ideas and new perspectives to America’s foreign policy,” according to Secretary of State Colin Powell.

The researchers were announced as the first Jefferson science fellows on 26 May. The group, which includes an evolutionary biologist, a biochemist and a remote-imaging specialist, will take up their posts in August.

The Jefferson Science Fellows programme allows tenured researchers to work in Washington DC for a year, advising policy-makers on science-related issues. In return for their year’s education in statecraft and diplomacy, the researchers must act as

consultants for the state department for five years after they return to their home institutions.

India installs former lawyer as science minister

New Delhi India's new coalition government has named Harvard-educated lawyer turned politician Kapil Sibal as science minister.

Unlike his predecessor, Murli Manohar Joshi, Sibal will not serve as a full cabinet minister. He will hold the lower rank of minister of state, and will have to seek cabinet approval on policy issues. But science secretary Valangiman Ramamurthi says this should not be seen as a signal that science has been given lower priority by the new prime minister, Manmohan Singh.

Sibal's appointment has been generally welcomed by India's scientific community, despite his lack of science training. He has promised to simplify regulations in biotechnology and to give top priority to stemming India's brain drain.

Stem-cell research in line for cash bonus

London The UK research councils have given stem-cell research a £16.5-million (US\$30-million) boost. The funding will

NASA releases wreckage for shuttle study

Washington NASA has loaned several pieces of the wreckage of the space shuttle Columbia to an independent aerospace researcher.

The move marks something of a departure for NASA. The remnants of two previously destroyed spacecraft — the shuttle Challenger and the burned Apollo 1 capsule — were stored away out of respect for the lost astronauts. But last year, following a recommendation by shuttle launch director Mike Leinbach (right), NASA decided to grant companies access to the Columbia debris for research.

The first recipient, materials scientist Gary Steckel of The Aerospace Corporation in El Segundo, California, will use propellant tanks and one of the shuttle's engines to study how the heat and pressure of atmospheric



re-entry affects the composite materials used in such parts. NASA notified the families of the Columbia astronauts before releasing the debris.

go towards 57 research grants in embryonic, adult and fetal stem-cell research, including the development of new embryonic stem-cell lines.

This should be welcome news to the country's stem-cell bank at the National Institute for Biological Standards and Control in Potter's Bar, which opened last month with just two cell lines. Britain has much more liberal rules regarding stem cells than countries such as the United States, something that many think should lead to a

research advantage. "This is the first direct attempt to exploit our lead in stem-cell research," says Richard Oreffo from the University of Southampton's School of Medicine.

"Without a boost like this, the path of stem-cell research would be long and tortuous," says Ann Logan, a molecular neuroscientist at Birmingham University who has received £400,000 to study how adult stem cells may be used to regenerate brain neurons.

Running before we can walk?

Two years ago, a new proteomic test was heralded as the future of cancer diagnostics. But since then, doubts about its effectiveness have begun to grow. Erika Check reports.

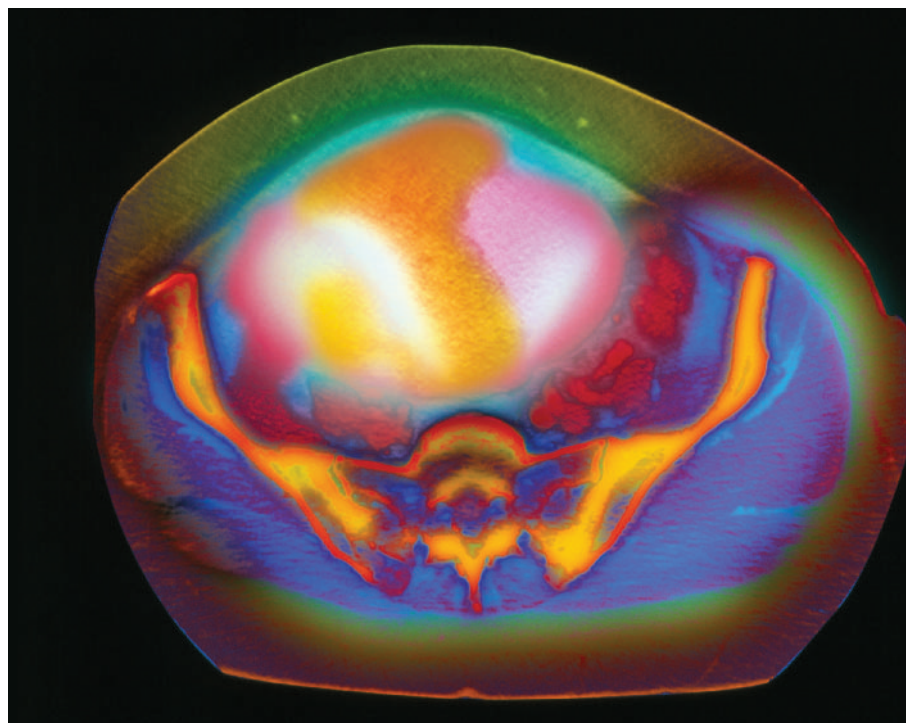
Seldom does a single piece of research prompt the US Congress to pass a resolution urging continued funding to drive a new diagnostic test towards the clinic. But that's what happened in 2002, when *The Lancet* published a paper¹ claiming a breakthrough in the diagnosis of ovarian cancer.

The paper described the use of mass spectrometry to analyse the pattern of proteins present in samples of blood serum. On the basis of these patterns, the test detected all the patients with ovarian cancers in a set of 50 samples, and falsely identified just three healthy patients as suffering from the disease from a total of 66 control samples.

Most encouragingly, the technique seemed to work well on patients with early-stage cancer — offering the prospect of earlier diagnosis, which improves the chances of successful treatment. The best current blood test, which relies on the detection of a single protein called CA125, misses at least half of patients in the earliest stages of the disease, and gives a high rate of false positives.

The researchers, led by Lance Liotta and Emanuel Petricoin, who co-direct a proteomics programme run by the National Cancer Institute and the Food and Drug Administration, based in Bethesda, Maryland, won immediate acclaim. In addition to the congressional resolution urging further funding for their research, the consumer magazine *Health* named the test one of the top ten medical advances of the year.

Commercial rights to develop the test were quickly licensed from the US government to a company called Correllogic Systems, also based in Bethesda, whose scientists collaborated with Liotta and Petricoin on the



On target: can proteins in the blood reveal ovarian tumours (pink/yellow) before they reach this stage?

Lancet paper. In November 2002, Correllogic granted licences to two larger firms, Quest Diagnostics and the Laboratory Corporation of America, which are now hoping to market the test under the brand name OvaCheck.

But those plans could be thrown off track by reanalyses of Liotta and Petricoin's data by independent groups, which have raised serious doubts about OvaCheck's reliability.

These questions prompted the Society of Gynecologic Oncologists to review all of the published work about OvaCheck. On 7 February, the society declared that "more research is needed to validate the test's effectiveness before offering it to the public".

Early warning

Critics warn that the episode illustrates the dangers of moving rapidly to the clinic with immature technologies such as those of proteomics. They say that scientists and regulators need to develop standards to ensure that such tests really work before they hit the market, because early detection is not without risks. Women who get false positive results may undergo unnecessary surgery, and those who get false negatives may forgo further screening. "Early detection is not a benign undertaking," says Martin McIntosh, who runs a cancer-detection effort at the Fred Hutchinson Cancer Research Center in Seattle, Washington.

The first criticisms of OvaCheck hit the public domain in June 2003, when two biostatisticians at the University of Maryland in Baltimore, James Sorace and Min Zhan, published a paper in the online journal *BMC*

*Bioinformatics*². They had reanalysed a data set that Liotta and Petricoin's team posted online in August 2002. Sorace and Zhan similarly found numerous differences in the protein patterns that discriminated between the cancer patients and the healthy controls. The trouble, according to Sorace and Zahn, was that these looked more like experimental artefacts than real biological differences.

The proteomics test relies on using gravity and electric fields to separate the proteins in a given sample. Each protein is then given a number that represents the ratio of its charge and mass — called its *m/z* value. The test

identifies patterns in these numbers to give a diagnosis. Sorace and Zahn were concerned because the most marked differences between the cancer patients and unaffected controls occurred for proteins with *m/z* values of less than 500. Many spectrometry experts consider *m/z* values below 2,000

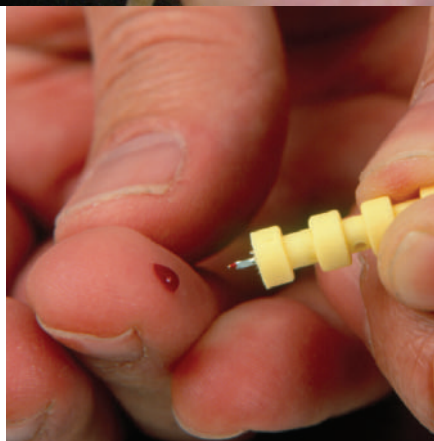
to be suspect, because they tend to include values generated by experimental artefacts or measurement error. To Sorace and Zahn, this suggested that the data for cancer patients and controls had been collected under such different conditions that it was impossible to reliably identify true biological differences.

Further questions emerged when biostatisticians led by Keith Baggerly of the MD Anderson Cancer Center in Houston, Texas, reanalysed the data in the *Lancet* paper, plus two further data sets posted online. "Based on these data, we really can't tell if it's possible to use proteomics to separate normals from cancers," Baggerly says.

"Whether or not OvaCheck works, we will learn from this experience what rules of evidence we might apply in the future to find useful results more efficiently."



Emanuel Petricoin (above) holds a protein pattern generated by the blood test he believes can reliably diagnose cancer.



Baggerly's first concern was that the values posted on the proteomics programme's website had already been processed in a way that made it impossible to reconstruct the raw data. But more specific concerns arose when his team analysed the overall characteristics of the data, looking at how closely the peaks for each of the proteins in the samples matched one another. Each of the data sets was divided into three groups: cancer patients, unaffected controls and patients with benign tumours. In the first data set, the pattern of proteins detected in the benign-tumour group seemed very different from the patterns detected in both the cancer and normal groups. In contrast, the patterns generated by the cancer and normal groups looked highly similar to each other.

Furthermore, the patterns detected in the first benign-tumour group looked almost identical to the patterns detected in all three groups in the second data set — cancers, normals and patients with benign tumours. This indicated to Baggerly that Liotta and his colleagues inadvertently changed their

experimental set-up midway through collecting the first data set — the one that they analysed for their *Lancet* paper.

Although the second and third data sets seemed consistent with each other, the precise differences in protein patterns that separated cancers from controls in one set could not be used to make the same distinction in the other — which makes little sense if the test is uncovering fundamental biological differences between cancerous and control samples. Baggerly's team also claimed that the data were collected on a machine that had not been properly calibrated, or adjusted for accuracy. In a paper first published online in January this year³, the researchers concluded that the test may be uncovering differences due to "artifacts of sample processing, not to the underlying biology of cancer".

Divided over data

Baggerly says that he discussed his concerns privately with Liotta and Petricoin in December 2002. He decided to publish only after learning that a test might become commercially available.

Petricoin says there is no proof that m/z values below 2,000 always represent experimental noise or bias. He also denies that the team switched methods midway through the first experiment. And he explains that the second and third data sets were processed differently. "The point is, the *Lancet* paper shows feasibility of this approach, and the results derived from each of these data sets prove that there do indeed exist low-molecular-weight molecules in the circulation that can

discriminate the disease states," says Petricoin, who adds that the team has further refined its methods in a new paper⁴. He also notes that other groups have examined his data and supported his conclusions⁵.

Petricoin stresses that he and Liotta are not directly involved with commercial development of the OvaCheck test. Peter Levine, Correllogic's chief executive, says that the company has also refined its data analysis techniques since the *Lancet* paper. "It seems to me a lot of people are sort of debating an issue that is pretty much of historical relevance only," he says. But Correllogic has not released its data, so it is impossible to verify this claim.

Meanwhile, Correllogic is now in dispute with Liotta and Petricoin over the pair's consulting work for a rival company, Biospect of South San Francisco. On 18 May, the two scientists were called before a congressional committee and asked whether this had slowed OvaCheck's development — a charge that they roundly denied.

Growing pains

But this dispute is secondary to the main issue of whether the technology works. Sceptics want to see more evidence that it yields consistent results on samples from different labs. And some argue that the field needs to develop standards for how proteomics experiments should be done and reported.

Because the technology is so new, says epidemiologist David Ransohoff of the University of North Carolina in Chapel Hill, scientists are still learning how to cut out all the possible sources of bias in proteomics experiments. For instance, if the cancer samples are collected from women being tested because they are known to be at high risk of suffering from the disease, they might experience anxiety when sampled, unlike controls whose samples may be taken as part of a routine check-up. In this case, the first group of samples may be flooded with stress hormones that would be detected by a proteomic analysis, but have nothing to do with whether or not a woman has cancer.

"Whether the test works or not, we will learn from this experience with OvaCheck what rules of evidence we might apply in the future to find useful results more efficiently," says Ransohoff.

He hopes that the episode will lead to an established set of standards for evaluating the effectiveness of such tests. Only then will we know whether proteomics-based diagnostic tools truly deserve the trust of scientists, doctors — and, most importantly, patients. ■

Erika Check is Nature's Washington biomedical correspondent.

1. Petricoin, E. F. & et al. *Lancet* **359**, 572–577 (2002).
2. Sorace J. M. & Zhan, M. *BMC Bioinformatics* **4**, 24 (2003).
3. Baggerly, K. A., Morris, J. S. & Coombes, K. R. *Bioinformatics* **20**, 777–785 (2004).
4. Conrads, T. P. et al. *Endocr.-Relat. Cancer* **11**, 163–178 (2004).
5. Zhu, W. et al. *Proc. Natl Acad. Sci. USA* **100**, 14666–14671 (2003).

► <http://ncifdaproteomics.com>



One of a kind

Lonesome George is probably the last giant tortoise of his type. But are scientists doing all they can to find him a partner, boost his sex drive and save his subspecies? Henry Nicholls finds out.

Poor Lonesome George. He may be famous, but he hasn't got a mate. All alone in the world and, sadly, singularly uninterested in sex, George the Galapagos giant tortoise looks set to be the last in his line.

Ever since George was discovered in 1971, there have been many attempts to get him to reproduce. Researchers have combed his island and the world's zoos in search of a female of the same subspecies. They have brought in a pair of female tortoises from a nearby island to act as playmates. They have even enlisted the help of a young, attractive female zoologist from Switzerland to help boost his sex drive.

So far they have met with little success: Lonesome George hasn't shown the slightest interest in the females in his pen. But his keepers do have other cards to play. George has yet to be introduced to what may prove his best possible partners — female tortoises from a distant island that seem to be more closely related to him than those currently in his pen. If researchers can get the two together and overcome George's low libido, he may yet reproduce.

The Galapagos Islands were once crawling

with giant tortoises. Before the first ship sailed into the Galapagos archipelago in 1535, it was home to at least 15 distinct populations of giant tortoise (*Geochelone nigra*), isolated from each other on different islands or on volcanoes separated by impassable lava flows. But heavy exploitation at the hands of eighteenth- and nineteenth-century whalers, along with competition from introduced animals such as goats, pigs and rats, caused four of these subspecies to go extinct and has earned the remaining 11 a place on the World Conservation Union's list of threatened species. Indeed, the *G. nigra abingdoni* subspecies from the island of Pinta was thought to be extinct until a snail biologist stumbled on a sole survivor: Lonesome George.

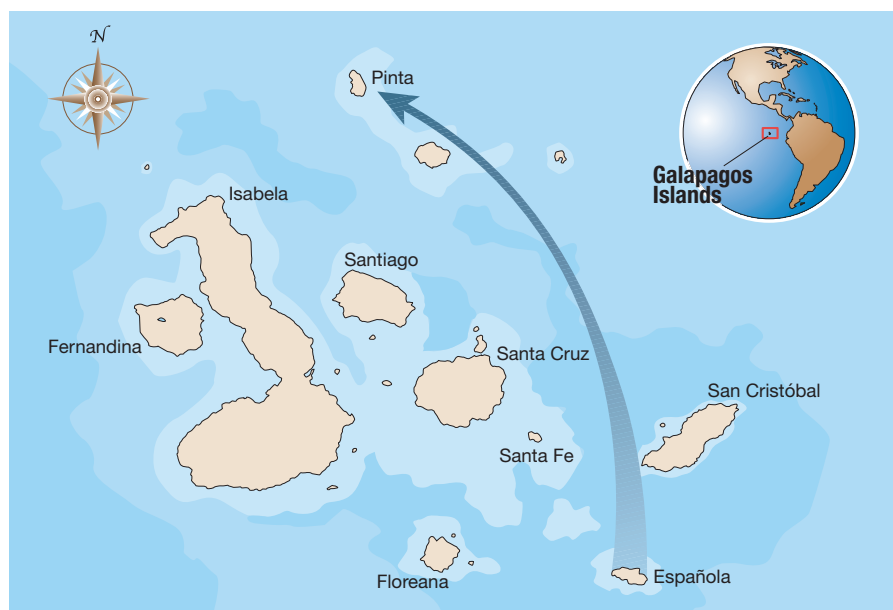
Lonely existence

George lives at the Charles Darwin Research Station (CDRS) in Puerto Ayora, a bustling tourism-driven town on the Galapagos island of Santa Cruz. There he serves as a mascot for the archipelago's conservation-based research. The ultimate goal of his scientific custodians is to restore the Pinta ecosystem, which means repopulating the

island with giant tortoises — ideally, with Lonesome George's descendants.

For many years, that was thought to be a lost cause. What could you do with a single male tortoise? "Conservation work in Galapagos reminds me of triage," says Linda Cayot, former head of protection at the CDRS. "The cases that look to be beyond saving are set aside to deal with those that are great emergencies but where the patient can be saved." For 20 years George was left alone, until, in 1992, Cayot put a few females from the nearby island of Isabela in his pen, hoping that romance would blossom. At the time it was reasonably assumed that tortoises from this island — the nearest geographical neighbour to Pinta that has tortoises on it — would be most closely related to him and so offer the greatest chance of a successful sexual liaison.

But the first examination of genetic similarities between the different tortoise populations in the Galapagos, done in the late 1990s, threw up something of a surprise. DNA analyses conducted by Gisella Caccone, an evolutionary geneticist at Yale University in New Haven, Connecticut, and her colleagues showed that Lonesome George is more closely related to the subspecies on the island



Distant relatives: genetic analyses suggest that George's ancestors travelled from Española to Pinta.

of Española (*G. nigra hoodensis*). This indicated that giant tortoises, which are ill-suited to life at sea, somehow survived an incredible 300-kilometre journey from Española in the south to Pinta in the north (see Map), probably by hitching a ride on strong currents.

More than a scientific curiosity, the result means that George has not been paired up with the most appropriate females. "Now that we see he has close genetic affinities to the Española and San Cristóbal subspecies, perhaps they would be a more appropriate source of a mate for this sole survivor," concluded the authors in their paper. But five years on, their recommendation has not been acted on.

Unnatural selection

This is because the Darwin station abides by an ethical principle outlined by the Galapagos National Park Service — trying to ensure that biological processes continue as if humans had never set foot on the islands. Encouraging Lonesome George to mate with a tortoise from another island is an unnatural intervention that goes against this mandate, says Howard Snell, director of science at the CDRS.

This view is clearly at odds with Cayot's attempt at matchmaking between the Isabela females and George. But there has been no move to split them up, probably because there seems to be little need: in the dozen years that they have been together, Lonesome George has shown a spectacular indifference towards them.

Snell's attitude may sound odd, as it seems to favour the total extinction of a species over any 'unnatural' intervention that might preserve at least some of its genes. But Snell isn't completely opposed to cross-breeding. He is just biding his time, hoping to come across a better solution before having to

resort to contaminating George's gene pool.

Snell's first priority is to find a living Pinta female for George. This, though unlikely, isn't impossible. In spite of their size, giant tortoises can be very hard to spot. Two tortoises live on the tiny island of Santa Fe, says Snell, which is visited by at least 50,000 tourists a year and serviced by three research teams. Yet these tortoises are seen only once every seven or eight years.

And there has been at least one tantalizing hint of tortoise life on Pinta. Back in 1981, nearly ten years after Lonesome George had left the island, Cayot and a colleague came across a lump of tortoise excrement near the top of Pinta's volcano. "As I stepped across a grassy area I looked down and saw a tortoise scat," she recalls. "I yelled out and we all stopped. We were all amazed." The scat hadn't completely decomposed, stirring hope that the faecal specimen had come from another individual still roaming the island.

But so far no one has seen it. Late last year, Peter Pritchard, founder of a privately funded conservation group in Florida called the Chelonian Research Institute, gathered together more than 20 park staff and conducted the most thorough search of the island to date. They had mixed success — they found a total of 15 Pinta tortoises, although all of them were dead and only one of them was female.

Singles club

There is a chance, however, that there may be Pinta females elsewhere, in private collections or zoos. Edward Louis, a conservation geneticist at Henry Doorly Zoo in Omaha, Nebraska, has collected DNA samples from more than 400 Galapagos tortoises from zoos worldwide. None looks particularly like a Pinta tortoise — although with only Lonesome George's genes to go

on, it is impossible to be sure what actually constitutes a Pinta tortoise. Still, there are more than 100 zoo-based tortoises yet to sample, estimates Louis. "There are a lot of animals in South America out there that we haven't even looked at," he says. The Quito Zoo in Ecuador has "a tonne", he notes.

The CDRS itself is home to 23 adult giant tortoises of uncertain geographical origin, mostly garnered over the years from private collections. Some have been in captivity since the 1960s, and until 1976 (before the purist ideology came into effect) were allowed to mate freely with each other. Between them, they have produced a rabble of 36 offspring — most the result of undocumented pairings. Caccone and her colleagues have used DNA sequences to work out the geographical origins of the adult tortoises and the genetic origins of the progeny they produced². Sadly for Lonesome George, none of them seems to be a Pinta tortoise.

Long-distance romance

Most of the giant tortoises in captivity are not held in research stations or in zoos but are in private collections, says Snell. But these are much harder to track down. The permit requirements for owning a Galapagos tortoise are very, very confusing, and people are probably unsure whether their animals are legal, Snell explains. What's needed is a worldwide amnesty that would allow this population to be quantified and sampled, he adds. But such a move is logistically fraught.

As each day passes without news of a Pinta female, the CDRS comes ever closer to compromising its purist approach and turning its thoughts towards hybridization. If the work of geneticists is heeded, the Isabela females should be replaced by a closer relative from Española.

Although the Española population is itself under threat of extinction, things are far healthier than they are on Pinta. In the 1960s, the population fell to just 14 tortoises — 12 females and 2 males — that were brought into captivity at the CDRS. A third Española male was subsequently shipped in from the San Diego Zoo to be part of a captive breeding programme that aims, one day, to restore the Española subspecies to something like its former glory. Since 1975, CDRS scientists have returned more than 1,200 tortoises to Española. More than 15 years ago, nests once again began to appear on the island.

But things may not be as healthy as they seem. Michel Milinkovitch, head of the evolutionary genetics unit at the Free University of Brussels in Belgium, is part of a team that includes Snell and Caccone that has been looking at the genetics of this population³. Its analysis indicates that not all of the tortoises are contributing equally to the next generation. The male brought in from San Diego,



Take your partner: most male tortoises are keen to acquaint themselves with the opposite sex (right), but Lonesome George (top left) is less accommodating.

known as Macho because of his sexual prowess, seems to have sired more than 60% of the offspring. Such a skewed genetic contribution “raises the spectre of severe inbreeding”, warns Milinkovitch.

On the positive side, the sheer number of tortoises to have come out of the Española programme means that there should be plenty of sexually mature Española females to pair up with Lonesome George, says Milinkovitch. There are even four Española females already housed in George’s neighbourhood at the CDRS, according to Caccone’s survey of its tortoises of unknown origin.

No sex, thanks...

Whoever his keepers decide to give George as a partner, there is still the problem that he seems to have no interest in sex. If this can’t be solved, then all this talk of reproduction will remain purely academic.

Normally, a male giant tortoise will mate with almost anything. “It will mate with a dead tortoise; it will mate with a rounded rock on its trail if it’s in the mood,” says Pritchard. “But Lonesome George seems to be cut from a different cloth.”

One possible explanation for George’s indifference is that he hasn’t been exposed to normal mating or courtship behaviour. Introducing a competitor male into his pen might goad him into activity, says Pritchard. “It might also be good if George was able to watch the active courtship and mating of Española tortoises in the adjacent enclosure,” he says. If none of these measures succeeds in getting George going on his own, then artificial insemination might be the only option.

Sadly, nobody actually knows if Lonesome George can produce any sperm. In

1993, only a year after the Isabela females had joined him in his pen and while the disappointment at his inactivity was still freshly felt, German vet Gisela von Hegel came to the Galapagos, and showed the staff how zoo-keepers usually examine the sexual health of a male tortoise. If the animal’s rear end is lifted off the ground, its muscles relax and its penis will drop out of the cloaca. This can then be manually stimulated.

Shortly afterwards a young and attractive female zoology student from Switzerland, Sveva Grigioni, arrived at the CDRS and volunteered her services. Initially things looked promising. “Sveva could get the other male tortoises in the exhibition pens to produce sperm within 15 minutes of approaching them,” says Cayot. But with bashful George, things did not move so fast.

An intimate relationship

Grigioni spent several hours a day with him, gaining his confidence and a nickname — Lonesome George’s girlfriend. “Day by day, he started to be more interested in the females with him,” says Grigioni. But after nearly four months of daily visits to his enclosure, Grigioni’s visa ran out and her work with George came to an abrupt end. “If I had had more time, perhaps I would have had more success,” she says. “I am almost certain that I would have got sperm.”

But she didn’t. In the absence of a very patient researcher, there are other, harsher methods that might induce George to come up with the goods. Giving an animal a small electric shock does tend, in most instances, to generate a response, and is commonly used in many captive breeding programmes to obtain sperm⁴. But it is hardly a non-invasive

procedure. Nobody knows how Lonesome George would react to having a probe inserted into his cloaca and an electrical current passed across his testes. So far, no one has been willing to find out.

If Lonesome George is infertile, what other options are there for the survival of his subspecies? The answer, according to the tourist information panels that surround his enclosure, is the fantastic world of cloning. But if there is one thing that the scientists with an interest in Lonesome George all agree on, it’s that there are other problems in the Galapagos on which money would be much better spent.

“Cloning George is ridiculous,” says Caccone. “It was so difficult with Dolly the sheep and we knew so much about mammals. The task is daunting for reptiles.” The effort required to clone each new species is huge, admits Dolly’s cloner Ian Wilmut of the Roslin Institute near Edinburgh, UK, but cells from George and the other species of Galapagos tortoise should be stored anyway. “Who knows, they might be used for cloning at some distant time in the future,” he says. As yet, cells have not been collected.

The one thing that Lonesome George has got going for him is time. He is probably less than 100 years old — young and sprightly for a giant tortoise — and might live for another 100 years. With patience and luck, George may yet be a father.

Henry Nicholls is a freelance writer based in London.

1. Caccone, A., Gibbs, J. P., Ketmaier, V., Suatoni, E. & Powell, J. R. *Proc. Natl Acad. Sci. USA* **96**, 13223–13228 (1999).
2. Burns, C. E. *et al. Anim. Conserv.* **6**, 329–337 (2003).
3. Milinkovitch, M. C. *et al. Proc. R. Soc. Lond. B* **271**, 341–345 (2004).
4. Silva, A. R., Morato, R. G. & Silva, L. D. M. *Anim. Reprod. Sci.* **81**, 159–175 (2004).

♦ www.galapagospark.org/en/home.htm

Cultural weight dragging at Asian giants' feet

Scientific progress is impeded by a tradition of conformity and respect for authority.

Sir—Reading the Commentary “Cultural reflections” by Mu-ming Poo, on the subject of Chinese science (*Nature* **428**, 204; 2004), I was struck by the similarities between science in China and in India.

Poo concludes that mediocrity in Chinese science should be attributed to cultural factors such as conformity and respect for authority, rather than to pure economics. India, like other Asian countries, is at a crossroads. It is time for these countries to free themselves from the burden of their past and to develop a modern structural framework for fostering original scientific research.

Both India and China are emerging into modernity. Each has a booming urban economy (with vast problems in the rural areas) and India's gross domestic product (GDP) has suddenly increased in recent years, thanks to liberalization and private entrepreneurship. India's overall funding in basic research has also improved, and now it is 1.2% of GDP.

Yet although India has a considerable number of scientists, it cannot be counted as a major world player in basic sciences,

notwithstanding its performance in space and nuclear technologies.

Not a single recent issue of *Current Science*, India's premier science journal, has appeared without at least one critique of India's performance in basic sciences, for example concluding that India and China face very similar problems (P. Balaram *Curr. Sci.* **86**, 755–756; 2004).

Twenty years ago, John Maddox wrote, about science in India, “Among developing nations, India has by far the best chance of succeeding” (*Nature* **308**, 581–584; 1984). However, he also wrote, “India has set ambitious goals for science and technology — self-reliance and the relief of poverty. Some great things have been accomplished, but much effort is frustrated.”

The Pakistani physicist and Nobel laureate Abdus Salam has been quoted as saying during a visit to India in 1981, “Not a drop ... has been added by India and other developing countries to ‘the pool of world knowledge’.” (H. Narain *Curr. Sci.* **65**, 739–742; 1993).

I believe that nothing much has changed during the intervening years.

Why cannot India and China, two Asian giants, break the mediocrity barrier, despite their strong fundamentals?

Culture is a major impediment, as Poo points out. Although India is a democracy, the Asian tradition of respecting authority and hierarchy remains strong in our society, spawning conformity and nepotism. Creativity in science can be fostered only in a free and unfettered intellectual environment.

Cultural shifts take time. Meanwhile, I believe structural changes can help these fast-developing Asian countries to achieve more in terms of high-impact scientific research, whatever the level of investment.

Starting with universities and other centres of higher learning, such measures should include introducing innovative courses in basic sciences at post-graduate levels (in India, the institutes of technology are a good place to start) and fostering better links between research organizations and industries.

C. P. Rajendran

Centre for Earth Science Studies, Akkulam, Trivandrum 695031, India

Ecological and political costs of river diversion

Sir—In the Commentary “Agriculture of the future” (*Nature* **428**, 215; 2004), T. C. Tso discusses the problems faced by China's agricultural sector and suggests several possible approaches to the problem of feeding the world's largest population. We support many of the suggested solutions, such as reducing farmers' costs, providing more funds for agricultural research and adopting new techniques to improve agricultural products.

However, we strongly oppose Tso's proposal to divert water from southwest China to its northern and eastern areas, especially the idea of diverting water from the ‘big U-turn’ of the Yarlung Zangbo river in Tibet.

The ‘big U-turn’ is located in the geologically fragile region of the Himalaya mountains where the Indian Plate meets the Eurasian Plate. The problems of geologic instability and frequent landslides alone are enough to make any huge hydraulic projects difficult or even impossible. Additionally, the Yarlung Zangbo is an international river, passing through China, India and Bangladesh. A diversion project will undoubtedly change the hydrological conditions, especially the

volume of water available to the lower countries, and is likely to cause an international dispute. Moreover, the ‘big U-turn’ is located in one of the world's biodiversity hotspots. Tso's proposed water-diversion project would irreversibly damage the ecosystems of the world's deepest and longest canyon, particularly the aquatic fauna.

Finally, we believe that China's water problems would be best addressed by the adoption of sound water planning and management, water conservation and efficient irrigation techniques — not by large-scale diversions.

Anping Chen*, Changdu Chen†

**Department of Ecology and Evolutionary Biology, Princeton University, Princeton, New Jersey 08544-1003, USA*

†Department of Ecology, Peking University, Beijing 100871, China

named by *Time* magazine as Person of the Century, Albert Einstein (*Time* **154**, 27; 1999). Casti suggests that Turing's 1936 paper provided the “theoretical backbone” for all computers to come.

Although Turing, a hero of mine, certainly was one of the greatest, we should keep in mind that his paper essentially just elegantly rephrased Kurt Gödel's 1931 results and Alonzo Church's extension thereof. It did not have any impact on the construction of the first working program-controlled computer. That was made in Berlin by Konrad Zuse in 1935–1941 and was driven by practical considerations, not theoretical ones.

In fact, the greatest impact that Alan Turing made on daily life was probably through his contribution to cracking the Enigma code, used by the German military during the Second World War, which is sometimes cited as a decisive event of the war.

Jürgen Schmidhuber

Dalle Molle Institute for Artificial Intelligence (IDSIA), Galleria 2, 6928 Manno-Lugano, Switzerland

Turing's war work counts for more than computers

Sir—John L. Casti, in his fine review of *Alan Turing: Life and Legacy of a Great Thinker*, edited by Christof Teuscher (“Touring artificial minds” *Nature* **428**, 258; 2004), proposes that Turing had more impact on everyday life than the man

correspondence

Contributions to Correspondence may be submitted to corres@nature.com. They should be no longer than 500 words, and ideally shorter. Published contributions are edited.

Women in the cabinet

The role of women in the history of science has long been overlooked.

Pandora's Breeches: Women, Science and Power in the Enlightenment

by Patricia Fara

Pimlico: 2004. 288 pp. £12.50

Jessica Riskin

Historians of science have been expanding their cast of characters beyond European male élites to include women, non-Europeans and members of the less-privileged classes. A second recent tendency has been to consider scientific ideas as inextricable parts of the social and cultural world in which they are generated. *Pandora's Breeches* gracefully straddles both movements. Patricia Fara's tactic in this elegant book is to use each of these approaches in the service of the other. Following female protagonists into the world of science during the Enlightenment, Fara enters its social and institutional dimensions, engaging in interpretation, translation and popularization. By focusing on these collective and communicative aspects of scientific work, she uncovers the main areas in which women were central to the pursuit of science.

Beyond their social and institutional roles, women occupied another niche, too. Early modern science involved much collecting and classifying, for which female assistants were often responsible. *Pandora's Breeches* takes a theme from this phenomenon, presenting itself as a sort of cabinet, a collection of women involved in the sciences. To the taxonomically inclined, a collection irresistibly invites classification, so here is my proposed taxonomy of the book's protagonists.

First come the inspirers: the correspondents and interlocutors. They include Elisabeth of Bohemia, who challenged Descartes to explain how immaterial soul and physical body can interact. There is Elisabeth's niece, Sophie Charlotte, Queen of Prussia, who pressed Leibniz on the problem of evil. And we have Anne Conway, whose theory of the intermingling of spirit and matter, worked out in correspondence with the Cambridge philosopher Henry More, helped shape Leibniz's ontology.

Next are the promulgators: the translators, commentators, illustrators and popularizers. Here we find Emilie du Châtelet, who translated Newton's *Principia* into French and wrote extensively on newtonian philosophy and its rivals, helping create a tradition of French newtonianism. There is Priscilla Wakefield, whose *Introduction to Botany*, which presented the linnaean system to women, became a standard text. And let us



Joining forces: Emilie du Châtelet spread Newton's ideas by translating *Principia* into French.

not forget Marie Paulze Lavoisier, who illustrated the chemistry texts of her husband, Antoine Lavoisier.

But Paulze Lavoisier belongs equally on several other shelves. For example, she should be among the assistants, having worked in Lavoisier's laboratory. Here she is accompanied by Elisabetha Hevelius, who conducted astronomical observations with her husband, Johannes Hevelius. The importance of her collaboration is signalled by the instruments they used, which included a sextant that required two people to operate it.

Finally, we have the institution-builders. Again we find Paulze Lavoisier, who hosted a weekly salon where people discussed scientific matters along with other gossip. She also oversaw Lavoisier's household. Fara compellingly argues that when experimentation took place in private houses, the wives who ran these houses played the role of laboratory managers. In this way, Jane Dee managed

John Dee's astrological and alchemical workshops and his staff of assistants.

How shall we classify Mary Shelley? Fara presents her as a commentator and critic, and views *Frankenstein* as a response to contemporary debates about the nature of life and a warning against the dangers of unbridled experimentalism. These are conspicuously peripheral roles compared with the other figures' scientific activities.

At the opposite end of the spectrum sits another member in a category of her own. Caroline Herschel conducted astronomical observations both with her brother, William Herschel, and also independently. They built massive, powerful telescopes and methodically charted the stars, an enormous labour of observation, calculation and classification. Caroline's own work yielded eight comets and several nebulae and star clusters. She published in the *Philosophical Transactions of the Royal Society* and gained

an international reputation, becoming an honorary member of the Royal Astronomical Society, winning various awards, and earning a salary from King George III. An astronomer in her own right, she stands out among the inspirers, promulgators, assistants and institution-builders.

The implications of any collection are best revealed by the entries that are hardest to classify. Shelley and Herschel, from opposite extremes, indicate the risks of including women by fundamentally transforming the story. Fara urges us to stop telling heroic tales of individual brilliance and show the sciences in their true light, as cooperative projects. To be sure, geniuses have never been 'lone'. But we risk confirming the traditional view that women are capable of collaboration but not brilliance. Does Jane Dee, who oversaw her husband's household, belong in the same collection as Emilie du Châtelet, one of only a handful of people in her generation capable of understanding Newton's *Principia*?

In other words, does this collection implicitly confirm that modern science is fundamentally a masculine philosophy and rationalism the attribute of masculinity? Or is there a different interpretation? Perusing this rich, beautifully crafted book you realize that, assuming that the sciences require both individual insight and community, even when women were continually assigned the latter, they often contributed the former. ■

Jessica Riskin is in the Department of History, Stanford University, Stanford, California 94305-2024, USA.

Big-game theory

The Kruger Experience: Ecology and Management of Savanna Heterogeneity

edited by Johan T. du Toit, Kevin H. Rogers & Harry C. Biggs
Island Press: 2003. 492 pp. \$75 (hbk), \$40 (pbk)

Andrew Illius

The Kruger National Park, a strip of bush 60 km wide that stretches 350 km from the tropical northern end of South Africa bordering Zimbabwe and Mozambique to the temperate south, is one of the world's great wildlife reserves. *The Kruger Experience* is not about the experience of being there, which is breathtaking, but about the accumulated experience of the managers and scientists who have worked for a century to conserve and understand it in all its glory. The editors faced the challenge of bringing together more than a hundred researchers and getting them to put their work into a common theoretical framework. The chosen 'cross-cutting theme' is ecological



Burns unit: park rangers use fire to manage the ecology of South Africa's Kruger National Park.

heterogeneity in time and space, and their intention is to unravel the truism that heterogeneity underlies biodiversity and adaptive management. But there is also considerable value in the main body of the work, which is a fascinating compendium of observations on the ecology and management of the savanna biome.

Given South Africa's turbulent political past, the history of Kruger's management can be viewed against the backdrop of the changing socio-political context. Most of Africa's great national parks were established at a time when land use was allocated to suit the economic interests of colonial settlers. The new South Africa takes more seriously the land-rights conflicts that arise between conservation priorities and displaced or neighbouring farming communities. There are now many examples, as in other parts of Africa, of community-based approaches that allow local people some control over wildlife resources in a way that is more beneficial to conservation and sustainable use than strategies that dispossess local people and reduce their access to such resources.

At the outset, management interventions in Kruger were somewhat crude, such as the control of predators and fire, and the establishment of a network of water points, and were intended to increase the opportunities for viewing game. These have now given way to more subtle forms of adaptive management that attempt to recognize the complexities of natural processes within defined conservation objectives.

Fire policy provides a good illustration of the evolution of adaptive management. Fire was initially regarded as something to be avoided, but has since been seen as an integral part of natural savanna systems. Fire suppression then gave way to a policy of prescribed rotational burning in 1956. This command-and-control approach was

abandoned in 1992, when it was realized that it had some negative effects and was a poor substitute for the processes by which natural fires drive and respond to vegetation heterogeneity. The next approach, which was intended to reproduce the patterns of frequency, season and intensity under which Kruger's biota evolved, allowed lightning fires to burn but suppressed anthropogenic fires. This led to the park's staff spending more time putting out fires than starting them. A mixed policy has now been adopted, in which all lightning fires are tolerated, and other fires are either started or tolerated only in areas that need to be burnt according to ecological criteria. A thorough-going state-dependent regime derived from the knowledge of how fires affect biodiversity was rejected as being too agricultural for a conservation area.

The ecological effects of establishing water points further illustrate the complexities of nature conservation and the law of unintended consequences. The provision of water was intended to allow game animals to spread into areas that were otherwise inaccessible to water-dependent species in the dry season, and to even out or homogenize vegetation use, but served to increase the number of wildebeest, zebra and their predators. Species that were less dependent on water, such as the roan antelope, now faced increased competition and heightened predation pressure from large carnivores attracted to the wildebeest and zebra. The roan antelope consequently suffered a precipitous decline in numbers, and is now endangered. In addition, the less dominant brown hyena has become extinct since the more dominant large carnivores became more common.

What, then, of heterogeneity? To those who reckon they have a common-sense understanding of heterogeneity, theoretical

treatments of it are something of an acquired taste, and the two fairly heavy-duty chapters here are no exception. Authors of subsequent chapters tend to focus on their specialism, and then say either how important it is for heterogeneity, or vice versa. A few chapters are on subjects sufficiently mature for the authors to really work the theme. For example, those on the ecology and population dynamics of the large herbivore community show how, over scales from biome down to local and from geological down to seasonal, habitat heterogeneity is a driving force in ungulate speciation and abundance, the size-scaling of ungulate assemblages, and the species richness of savannas. This and other examples show that understanding the effects of heterogeneity remains an important part of the research agenda for the future — a fitting tribute to the first century of research in Kruger. ■

Andrew Illius is at the School of Biological Sciences, University of Edinburgh, Michael Swann Building, Mayfield Road, Edinburgh EH9 3JR, UK.

More on South African environments South Africa's Environmental History: Cases & Comparisons

edited by Stephen Dovers, Ruth Edgecombe & Bill Guest

Ohio University Press/David Philip, \$24.95

Social History & African Environments

edited by William Beinart & JoAnn McGregor
Ohio University Press/James Currey, \$49.95 (hbk),
\$27.95 (pbk)

The Rise of Conservation in South Africa: Settlers, Livestock, and the Environment 1770-1950

by William Beinart
Oxford University Press, £65

Mind the gap

The Physiology of Truth: Neuroscience and Human Knowledge

by Jean-Pierre Changeux (transl. Malcolm DeBevoise)
Belknap Press: 2004. 288pp. \$45, £29.95,
€41.50

David Papineau

Can neurophysiology cast any light on the human condition? Books that set themselves this ambition, and there are plenty, are invariably disappointing. The problem is not that we lack information at the neuronal level — a great deal is known about cell receptors, neurotransmitters, re-entrant connections and so on. Rather, the difficulty lies in relating this microscopic knowledge to higher human faculties such as thought, emotion and consciousness.

Exhibition

Inspired by insects

Observation is key to the work of both artists and scientists. Illustration has been an essential part of scientific research for centuries and the images have often been admired for their aesthetic as well as informative qualities. Not surprisingly, the beauty of biology has also captured the imagination of artists, some of whom borrow scientific techniques as an aid to observation.

Artist Mark Fairington uses high-definition electron microscopes to photograph insect specimens before he begins to paint them. This enables him to capture on huge canvases minute details of the original, but with subtle manipulations.

Examples of his work, like the mantid depicted here, are on show in *Fabulous Beasts*, an exhibition at the Natural History Museum in London until September 2004. The exhibition also features the work of another artist, Giles Revell, who uses electron microscopy to reproduce majestic monochrome



images of common insects. The works of both artists are juxtaposed with specimens from the museum's own vast entomological collection, and a rare glimpse of the first edition of Robert Hooke's *Micrographia*.

To get round this, popular-science books by the likes of Francis Crick, Joseph LeDoux or Antonio Damasio typically have the following trajectory. We start with a few chapters on the neuronal nitty-gritty. But then the gears surreptitiously change, and we switch to speculation about the mind's higher powers. However, any serious theorizing at this level tends to be 'boxological', rather than physiological — we are given flowcharts connecting posited brain modules, but there is no bottom-up, cell-level account of how these modules might work.

Perhaps this is unsurprising, given the kind of evidence that is currently available about the large-scale operations of the mind. In recent years, functional-imaging data have been added to findings from studies of brain lesions. But even these new data are at too gross a scale: it is like trying to figure out how a computer works by noting when different bits get hot and what goes wrong when certain parts are broken. With luck, this might give us some idea of where certain operations are located, but it is not going to tell us about the mechanisms that make them possible.

Jean-Pierre Changeux's credentials as a neurophysiologist are outstanding. He has been director of the Unit for Molecular Biology at the Pasteur Institute in Paris for more than 30 years, where he has played a prominent role in understanding allosteric proteins and their relevance to neurotransmitter

reception. Nor is Changeux any stranger to popular science writing — his *Neuronal Man* (1983) and subsequent book-length dialogues with other prominent French intellectuals have been great successes in his native country and elsewhere. Nevertheless, his new book, *The Physiology of Truth*, suffers from the typical flaws of the genre. Initial chapters concentrate on neurophysiological signalling and modulation, but by the end the topics are knowledge, culture and the history of science. Interesting points are made at both levels, yet the initial neuronal material seems to shed little light on the later large-scale issues.

Still, the book does have the virtue of suggesting a possible, deeper explanation of why the micro-macro gap may be so hard to bridge. Throughout the book Changeux emphasizes the plasticity of the brain. Significant neuronal variation can be found among even simple organisms such as water fleas, and the brains of monozygotic human twins often exhibit striking differences. Changeux sees this variability as a result of epigenetic selection — the genes provide a general 'envelope' for brain development, but the details depend largely on the selective favouring of some spontaneously formed synaptic connections over others during development.

If this neural darwinism is right, then perhaps it is inevitable that any attempt to identify the neurophysiological mechanisms

Science in culture

A fluid definition of art

Viewing images of liquid crystals as art raises complex questions.

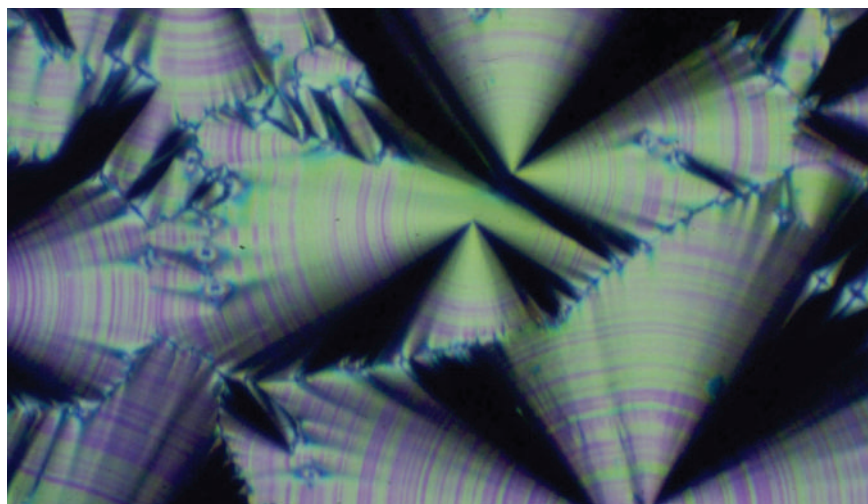
Martin Kemp

When does an image made by a researcher for scientific purposes become art — if at all? Over the ages, the beauty of scientific images has been widely recognized. Reading Robert Hooke's *Micrographia* (1665) gives us a sense of his aesthetic thrill faced with the wondrous images in the new device of the microscope. Like many of his contemporaries, Hooke would have recognized such miracles as the eye of the fly as the product of God's, or nature's, artistry. But does this mean that they are works of art in the normal sense of the term?

The question is implicitly raised by many of the striking products of modern scientific imaging techniques. And it is overtly posed by the claims of chemist John Goodby of the University of Hull, UK, that his microscopic images of liquid crystals (shown here) are "every bit as good as the kind of art you see in most galleries". Leaving aside the question of what is meant by "good" (and good for what?), his decision to start exhibiting his pictures as works of art plays into a complex series of shifting definitions of art in the modern era.

Until the twentieth century, the issue would not have arisen within the institutionalized definitions of art and science in post-Renaissance Western culture. But when artists decided to display everyday objects in art galleries — such as the signed urinal entitled *Fountain* by Marcel Duchamp in 1917 — the definition of art became very wobbly. The exhibiting of such 'ready-mades' and their enshrining in galleries and museums, in the collection of Duchamp's works in Philadelphia, for example, leaves us with a definition that extends little beyond the claim that anything is art that an artist claims is art — as is anything that viewers can look at as art.

Goodby's claims are of course more specific than merely saying that because he exhibits the liquid-crystal pictures as art, then they are art. He



is implicitly setting his images in the context of modernist abstraction, in which paintings or sculptures are devoid of figurative subject matter and narratives. Indeed, the way that the great masters of abstraction have transformed what we are prepared to consider as art has radically enhanced our ability to appreciate the marvellous natural configurations revealed by modern scientific techniques.

The amount of artistic contrivance in Goodby's images far exceeds that in Duchamp's urinal. The selection of certain liquid crystals at certain stages in their intermediate state between solid and liquid, the setting up of the microscope to deliver certain visual qualities, and the choices involved in rendering and printing the pictures (regarding colours, textures, plasticity, scale and framing, for example) are all done to create the best effect. This is to say nothing of the way Goodby collages his images to produce images of birds and flowers.

I wonder how many scientists who use visual images prominently in publishing their work have not made some kind of aesthetic choice at some time or other. Certainly anything that features on the cover of *Nature*, in its current format, is

designed to attract attention in ways that are comparable to the use of a painting on the cover of an art journal.

In the final analysis, should we worry about whether something is art or not? If it excites us, isn't that enough? My answer is drawn from a long historical perspective. The set definition of art as an aesthetic product devoid of practical function is actually comparatively recent (dating back to the late eighteenth century) and is limited to Western and Westernized societies. The art world has performed increasingly unconvincing conceptual gymnastics to accommodate everything that artists have recently thrown at it. If we stop being bothered by the question of whether something is art, and instead respond openly to the visual products that we are capable of making, we will be able both to agree with Goodby that his works are as 'good' as art, and say that any implicit competition between artists and scientist as makers of wonderful images is rather beside the point.

Martin Kemp is professor of the history of art at the University of Oxford, Oxford OX1 1PT, UK, and co-director of Wallace Kemp Artakt.

behind higher cognitive faculties will end in failure. Maybe there simply aren't any such mechanisms to be found. This is the view of 'functionalists' in cognitive science. They believe in large-scale patterns in human thinking of the kind portrayed in the familiar flowcharts, but they deny that there are any uniform physiological mechanisms to explain those regularities. Not that they assume any kind of spooky magic; rather, they argue that different mechanisms will underpin the regularities in different people.

From the functionalist point of view, asking about 'the physiological mechanism' responsible for scientific reasoning — to take a topic from the end of Changeux's book —

is like asking for 'the low-level explanation' of why all word-processing programs work in roughly the same way. In truth, there isn't any one such explanation. Different programmers use different tricks, subject only to the constraint that their programs end up doing what word-processors have to do. Similarly, neural darwinism may ensure that our brains use different tricks to achieve roughly the same ends, subject only to the constraint that we all end up getting around in the world reasonably well.

Changeux has plenty to say about neural darwinism, and touches on functionalism in passing, but he doesn't quite spell out the connection between them. Still, his book

presents a more satisfying picture of the brain than most of its competitors in this crowded market. On standard accounts, it can simply seem frustrating that we never get any bottom-up explanations of higher cognitive functions. If the structure of the brain is laid down by a definite genetic plan, then why can't we find out about the underlying mechanisms? Changeux's book fails to identify any such mechanisms too, but at least he gives us some insight into why the search for them may be doomed to permanent frustration. ■

David Papineau is professor of philosophy of science, King's College London, The Strand, London WC2R 2SL, UK.

When symmetry breaks down

Electroweak-symmetry breaking: solving the riddle of how symmetry is broken may determine the future direction of particle physics.

Edward Witten

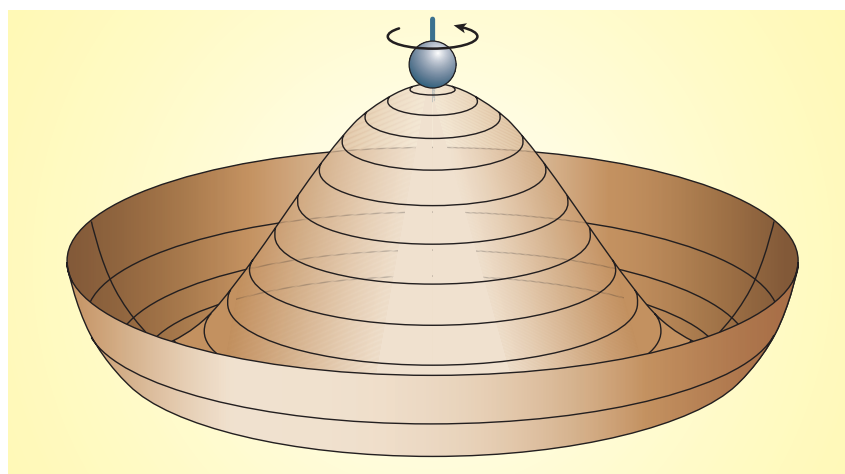
Our prehistoric ancestors did not need any modern equipment to detect the effects of what we now call the electromagnetic interactions. Light is pretty obvious in everyday life, and other electromagnetic effects, such as static electricity, lightning bolts and the magnetic properties of some rocks, such as lodestone, were well known in ancient days.

But it takes quite a bit of modern technology to discover even the existence of the weak interactions — let alone to understand them. We first became aware of weak interactions with the discovery of radioactivity in 1896. Some radioactive nuclei decay by emitting ' β -particles', which we now understand as energetic electrons. Such nuclear β -decay is the most accessible window into the weak interactions, and was the only one available until the middle of the last century, when cosmic ray detectors, nuclear reactors and particle accelerators arrived on the scene.

Electricity, magnetism and light once seemed to be three different topics. Their unified understanding, obtained in the nineteenth century, led scientists to describe them collectively as 'electromagnetic' phenomena.

Electromagnetism seems a good deal more obvious than the weak interactions. But in our modern understanding, which is based on something called the 'standard model' of particle physics, the two are perfectly in parallel. For example, electromagnetism is described by Maxwell's equations, and the weak interactions are described by a quite similar, albeit nonlinear, set of equations (called the Yang–Mills equations). To give another example, an elementary particle called the photon is the basic quantum unit of electromagnetism, and similar particles called W and Z bosons are the basic quantum units of weak interactions. Because of this close relationship between electromagnetic and weak interactions, modern particle physicists refer to them collectively as the 'electroweak' interactions.

If weak interactions are so similar to electromagnetism, why do they appear so different in everyday life? According to the standard model, the key is 'symmetry breaking'. Even if the laws of nature have a symmetry — in this case, the symmetry between electromagnetism and weak interactions, or between the photon and the W and Z bosons — the solutions of the equations may lack that symmetry.



Spoiling the spin: when the ball rolls down the slope, it will break the symmetry of the hat.

For example, in a liquid, an atom is equally likely to move in any direction in space — there are no preferred coordinate axes. But if we cool the liquid until it freezes, a crystal will form, which has distinguished axes. All directions in space are equally possible as crystal axes, but when the liquid freezes, some distinguished axes will always emerge. The symmetry between the different directions in space has been lost or 'spontaneously broken'.

Similarly, according to the standard model, just after the 'Big Bang' there was a perfect symmetry between the photon and the W and Z bosons. At the high temperatures that then existed, electromagnetism and the weak interactions were equally obvious. But as the Universe cooled, it underwent a phase transition, somewhat analogous to the freezing of a liquid, in which the symmetry was 'spontaneously broken'. The W and Z bosons gained masses, limiting the weak interactions to nuclear distances and putting their effects out of reach of the unaided eye. The photon remained massless, as a result of which electromagnetic effects can propagate over human-scale distances (and beyond) and are obvious in everyday life.

Most aspects of the standard model have been abundantly tested by experiment. For instance, the magnetic moment of the electron is measured down to the twelfth significant figure, with results that are in beautiful agreement with theory. Many predicted properties of the W and Z bosons have been verified to three or four digits. Most recently, the mechanism by which the standard model violates the symmetry between matter and antimatter has been tested in laboratories in California and Japan.

The one facet of the standard model that we have not yet been able to test experimentally is perhaps the most basic: how is the symmetry broken? However, we have a pretty clear idea of where such information can be found. Just as one can use atomic masses and binding energies to estimate the melting points of crystals, one can use the W and Z masses and other observed properties of elementary particles to estimate the high temperature or energy that particle accelerators need to achieve to explore electroweak-symmetry breaking. According to these estimates, electroweak-symmetry breaking may be in reach of the world's most powerful accelerator, the Tevatron at Fermilab in Chicago, and should certainly be in reach of the Large Hadron Collider (LHC), the new accelerator that is projected to go online in 2007 at CERN, the European particle-physics laboratory near Geneva.

What do we expect to find? In the original (and textbook) version of the standard model, the key to electroweak-symmetry breaking is an entity called the Higgs particle. At high temperatures, Higgs particles, like other particles, move at random. But as the Universe cools, Higgs particles combine into a 'Bose condensate', an ordered state in which many particles share the same quantum wave function, leading — in the case of helium — to superfluidity. The electroweak symmetry is broken by the 'direction' of the Bose condensate (in an abstract space that describes the different particle forces) in roughly the same way that in a crystal, the rotational symmetry is broken by the directions of the crystal axes. Although this proposal is simple and fits the known facts,

it is unlikely to be the whole story. A seemingly artificial adjustment of parameters is needed to make the Higgs particle mass small enough for the model to work.

Numerous proposed alternatives solve this particular problem, although they introduce puzzles of their own. One idea, motivated by a phenomenon that occurs in superconductors, is that the Higgs particle arises as a bound state. This would solve the problem of getting its mass right, but also requires a host of new particles and forces, which have not yet been observed. They should be detectable at the LHC. So far, models of this type have run into a great deal of difficulty, but maybe nature knows tricks that human model builders do not.

A more radical idea is 'supersymmetry', a new symmetry structure of elementary particles in which quantum variables are incorporated in the structure of space-time. The new symmetry prevents the particle interactions that would make the Higgs particle mass too big but, again, predicts a host of additional new particles that might be discovered at the LHC, and perhaps at the Tevatron.

Supersymmetry is one idea about electroweak-symmetry breaking that has had really convincing successes. A relation between different particle interaction rates based on supersymmetry is well confirmed experimentally. Moreover, our most interesting

attempts at a more complete unification of the forces of nature ('grand unified theories' and 'string theory') really only work if supersymmetry is assumed. On the other hand, supersymmetric models raise numerous perplexing questions to which human model builders do not yet have convincing answers. If supersymmetry is confirmed, then learning how nature dealt with those questions will probably give us crucial clues about a deeper understanding of nature.

Other ideas about electroweak-symmetry breaking go even further afield. One line of thought links this problem to extra dimensions of space-time, subnuclear in size, but observable at accelerators. This approach is probably a long shot, but the pay-off would be huge — discovering extra dimensions could give us the chance for direct experimental tests of the quantum nature of gravity and black holes.

Finally, another line of thought links electroweak-symmetry breaking to the dark energy of the Universe, which astronomers have discovered in the past few years by observing that the expansion of the Universe is accelerating. From this point of view, one tries to relate the relative smallness of the Higgs particle mass to the smallness of the dark energy. One approach is the anthropic principle, according to which the dark energy and the Higgs particle mass take different values in different parts of the

Universe, and we inevitably live in a region in which they are small enough to make life possible. If so, many other properties of the Universe that we usually consider fundamental — such as the mass and charge of the electron — are probably also environmental accidents. Although I hope that this line of thought is not correct, it will inevitably become more popular if experiment shows that electroweak-symmetry breaking is governed by the textbook standard model with a Higgs particle and nothing else.

As yet, none of these theoretical proposals about electroweak-symmetry breaking are entirely satisfying. Hopefully, by the end of this decade, experimental findings at the Tevatron and the LHC will set us on the right track. But the diversity and scope of ideas on electroweak-symmetry breaking suggests that the solution to this riddle will determine the future direction of particle physics. ■

Edward Witten is at the Institute for Advanced Study, School of Natural Sciences, Princeton, New Jersey 08540, USA.

FURTHER READING

Gunion, J. F. *et al.* *The Higgs Hunter's Guide* (Perseus Books, New York, 1990).
Kane, G. *Supersymmetry: Unveiling The Ultimate Laws Of Nature* (Perseus Books, New York, 2001).
Peskin, M. *Beyond The Standard Model* online at <http://arxiv.org/abs/hep-ph/9705479>

The atomic wrist-watch

Robert Wynands

Tired of having to set your watch every now and then? Well, how about an atomic clock on your wrist? Developments in technology and timing techniques could make this a distinct possibility.

Modern society relies heavily on precise timing. The most accurate timing devices — the primary atomic clocks in national metrology laboratories — each take up as much space as a large wardrobe and keep time with an accuracy of 1 part in 10^{15} , or 1 second in 30 million years. Not everyone needs this level of performance, and stability to about 1 part in 10^{12} is possible in shoe-box-sized clocks. Several thousand such devices are in use worldwide, for instance in telecommunications, to synchronize high-rate data transmission in multi-user networks. A much wider range of application would open up if the clocks could be miniaturized further and produced in an affordable way.

Solutions have now been found to some of the physical and technical problems encountered on the road to tiny atomic clocks. Liew *et al.*¹, in *Applied Physics Letters*, and in a conference contribution from the same group², describe the miniaturization of a vapour-cell atomic-frequency standard to a volume of just a few cubic millimetres, using a sequence of steps that are fully compatible with mass production. In *Physical Review Letters*, Jau *et al.*³ propose a novel mode of operation that could overcome some of the inherent disadvantages of small vapour-cell clocks. The result could be atomic clocks costing less than €100 (US\$120) each, whose size and power consumption would be suitable for hand-held devices, perhaps even for high-tech wrist-watches.

In a traditional atomic clock, atoms are illuminated by microwaves, in the gigahertz (10^9 Hz) frequency range. The microwave frequency can be regulated such that its absorption by an atom induces transitions between two specific internal atomic energy states at the highest rate possible. In this way, the regulated frequency of the microwave, the primary output signal of the atomic clock, becomes a macroscopic representation of the internal structure of the atom. In a caesium clock, the optimum frequency corresponds to

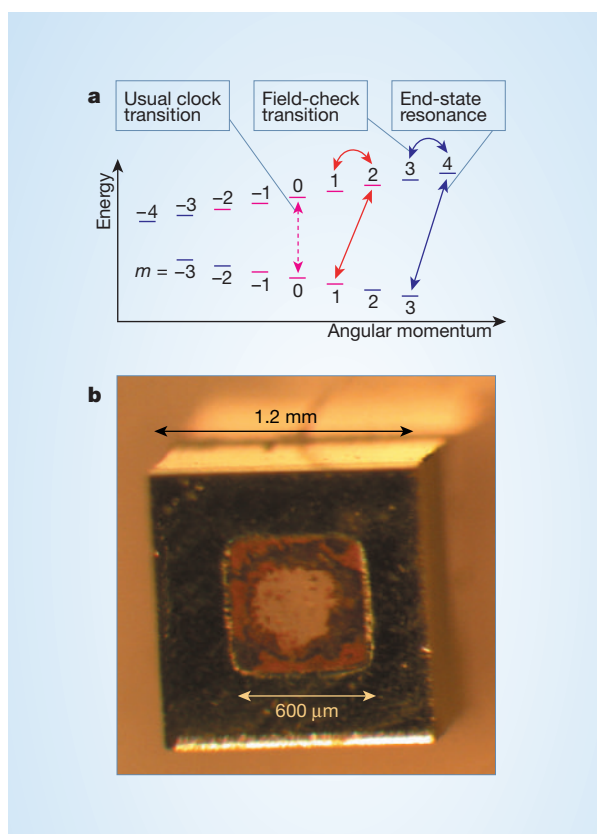


Figure 1 Clock mechanism. a, The energy levels of the ground state, split by the presence of a weak magnetic field, for the caesium atom (blue and red) and the rubidium atom (red only). The energy of each level is shifted in proportion to m , the magnetic quantum number related to the angular momentum of the atom. The usual clock transition is between the two levels with $m = 0$. Jau *et al.*³, however, propose that the end-state resonance (between $m = 3$ and 4 for caesium, and $m = 1$ and 2 for rubidium) is preferable for miniaturized atomic clocks. The strength of the magnetic field can be checked by monitoring the resonance frequency of the transition between neighbouring m states. b, A tiny caesium atomic-vapour cell built by Liew and colleagues¹. A square through-hole in a piece of silicon (0.375 mm thick) is filled with a small amount of caesium metal and then sealed off on each side with a thin glass sheet.

9,192,631,770 oscillations of the microwave per second (just over 9 GHz in frequency), so dividing the output electronically by this number provides one-second 'ticks' of the atomic clock. The specific pair of energy states usually chosen for clock operation are the central states within the 16-state manifold in the caesium ground state (Fig. 1a). For these states, their energy difference is only weakly perturbed by fluctuations in the magnetic field. In small clocks, the caesium atoms

themselves are typically contained in a glass cell of centimetre dimensions, held slightly above room temperature as an atomic vapour.

Miniaturization of such a microwave-driven clock cannot go below a centimetre or so because the microwave cannot fit into guide structures much smaller than its 3.2-cm wavelength. But this barrier can be broken by using laser light. When a light beam contains two optical-frequency components that are separated in frequency by 9 GHz, it can also couple the two atomic states, in a process called coherent population trapping. Finger-sized, all-optical clocks along these lines have been built before⁴, and Liew *et al.*¹ have now carried this work a crucial step further, by developing a new manufacturing technique for tiny vapour cells (Fig. 1b).

Their latest results² include the integration of the complete optical set-up (the 'physics package') in a device that takes up only a few cubic millimetres, not counting its electronics and power supply. Furthermore, all of the processing steps are fully compatible with existing technology for wafer-scale integration and processing. In principle, hundreds or thousands of physics packages could be manufactured simultaneously, keeping the manufacturing cost per unit low. The overall power consumption of the complete clock would be a few tens of milliwatts, so battery operation becomes a possibility.

But the miniaturization does come at the price of reduced stability: 2 parts in 10^{10} over 1 second, but almost 1 part in 10^{11} over longer times. The stability of the atomic clock depends on two parameters. It increases with increasing signal-to-noise ratio (SNR) during signal detection and decreases with increasing width of the microwave resonance line (that is, how much the 9-GHz frequency can differ from its optimum value).

In a miniature clock, the vapour cell is so short that the vapour density must be increased. To do this, the caesium atoms are heated to a temperature of more than 80 °C,

so that the same total number of atoms are producing a signal. At those high densities, the atoms frequently collide with each other. Each collision disturbs the atom's interaction with the laser light or the microwave in a process called spin exchange, which causes a broadening of the resonance line. Collisions can also transfer atoms away from the two energy levels involved in the clock transition, with the result that the effective number of atoms taking part in signal generation, and hence the SNR, is reduced.

Jau *et al.*³ propose to counter these effects by choosing a different pair of states, at the outer edge of the ground-state multiplet (Fig. 1a). So-called optical pumping, using circularly polarized laser light, can transfer almost all of the atoms into one or other of the outermost states, which leads to a large increase in signal strength and SNR. At the same time, spin exchange between atoms in one of the end states is not effective (because of conservation of angular momentum), so now spin-exchange collisions do not broaden the resonance line. Taken together, these two effects might well improve the clock's stability by a factor of ten to a hundred, taking the small clocks into the performance range — better than 1 in 10^{12} — of existing devices that cost about €50,000.

However, an important drawback of those end states is the strong dependence of their energy on the magnetic field's strength — which is why they are not normally used in atomic clocks with larger active volumes. Jau *et al.*³ propose to use a radio-frequency transition between neighbouring states to measure the magnetic field accurately in the small cell and then carefully compensate for any changes with an additional loop of electronics. The added complications and small uncertainties could well be tolerable in the face of the potentially large gain in short-term stability.

It will still be a while before a miniature atomic clock like that of Liew *et al.*¹ becomes a commercial product. Would people really want an atomic wrist-watch? An accuracy of 1 part in 10^{11} corresponds to a microsecond per day — perhaps a bit overblown for normal everyday use. But its small size and estimated low price will make it an important part of other products, such as improved GPS receivers or timing-dependent devices designed for autonomous operation. If the performance gains predicted by Jau *et al.*³ can be realized, we will then have affordable atomic time for the mass market.

Robert Wynands is at the Physikalisch-Technische Bundesanstalt, Bundesallee 100, 38116 Braunschweig, Germany.
e-mail: robert.wynands@ptb.de

1. Liew, L.-A. *et al.* *Appl. Phys. Lett.* **84**, 2694–2696 (2004).
2. Kitching, J. *et al.* in *Proc. 18th Eur. Frequency and Time Forum*, Guildford, UK, April 2004 (IEE, Stevenage, UK, in the press).
3. Jau, Y.-Y. *et al.* *Phys. Rev. Lett.* **92**, 110801 (2004).
4. Kitching, J., Knappe, S., Hollberg, L. & Wynands, R. *Electron. Lett.* **37**, 1449–1451 (2001).

Gene regulation

A reason for reading nonsense

Sabine Schmitt and Renato Paro

The process of transcribing DNA can itself regulate gene expression in yeast: in the case concerned it is the very act of reading the DNA, not the message produced, that carries out the regulatory job.

What is the production of DNA transcripts good for? Most notably, of course, the transcripts provide the read-out of an organism's genome in the form of messenger RNAs that act as blueprints for protein construction. Precisely defined sequence information is also required for structural or regulatory transcripts such as ribosomal RNAs, transfer RNAs or microRNAs. But the genome is transcribed at many more sites than is needed to produce these different RNA species. Why is there such a heavy traffic of RNA polymerases, the enzymes that copy DNA into RNA, and the production of large quantities of apparently non-coding and non-functional RNAs?

These RNAs are often dismissed as 'noise', perhaps caused by loose transcriptional control. But RNA polymerases are evidently doing more than we thought, and a new example is described on page 571 of this issue. Martens *et al.*¹ show how RNA polymerase II transcribes DNA sequences upstream of a gene and so controls the production of mRNAs from that gene. At the heart of this mechanism lies an intergenic

promoter, a DNA sequence that prompts the initiation of transcription that produces an RNA without protein-coding capacity. The authors find that processive transcription starts at this promoter and reads through the downstream promoter of the gene concerned, thereby shutting it off. The nonsense RNA transcript has no apparent function and is probably destroyed.

In the baker's yeast *Saccharomyces cerevisiae*, production of the enzyme SER3, involved in serine biosynthesis, is repressed during growth in a nutrient-rich medium. Surprisingly, Martens *et al.* found that when the SER3 gene is repressed, the upstream intergenic region is heavily transcribed. When they looked at the sequence of this region in more detail, they found conserved elements, one of which they identified as the promoter from which an intergenic RNA transcript originates. This transcript, termed SRG1 (for SER3 regulatory gene), does not itself show any significant sequence conservation and is devoid of open reading frames — nucleotide sequences that usually signal the presence of protein-coding capacity. The authors found that mutation of the

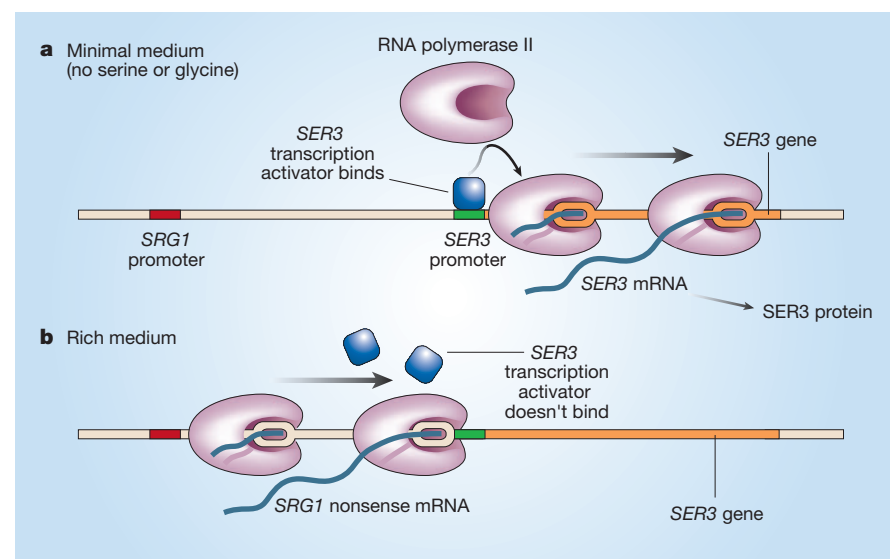


Figure 1 Intergenic transcription and gene regulation. a, During yeast growth on minimal medium lacking the amino acids serine and glycine, transcription of SER3 messenger RNA is required to produce the SER3 protein which is essential for serine biosynthesis. Here, SER3 transcriptional activator(s) bind to the SER3 promoter sequence, allowing RNA polymerase II to transcribe the gene. b, When yeast cells grow on rich medium, SER3 protein is not required. Repression of the SER3 gene is achieved by activation of the intergenic SRG1 promoter located upstream of SER3. After the initiation of transcription at SRG1, the RNA polymerase II machinery runs over the SER3 gene promoter, thereby preventing the binding of SER3 transcriptional activators and thus silencing the SER3 promoter¹.

intergenic promoter results in the disappearance of the transcript and in the lifting of *SER3* repression. They further show that the reciprocal expression pattern is not caused by the two promoter sequences competing for the same regulatory factors, or by the structural regulation of the downstream promoter by intergenic RNA.

But how, then, does the cell shut off transcription of a gene by transcription? Martens *et al.* found that if the *SRG1* transcript sequence is pared down to only 31 of about 550 bases, repression still occurs. But when *SRG1* transcription is terminated before the *SER3* promoter is reached, it does not. So it is actually the passage of the transcribing RNA polymerase machinery through the *SER3* promoter that leads to its shutdown by 'transcriptional interference'. In molecular terms, processive transcription prevents the binding of transcriptional activators to the *SER3* regulatory DNA elements and its promoter (Fig. 1).

Transcriptional interference has been demonstrated before, for example in yeast² and the fruitfly *Drosophila*³. What is new in the *SER3* context is that the process is governed by a precisely regulated intergenic promoter, and it illustrates a new facet of how non-coding sequences can influence gene expression. So far, most known gene-regulatory processes controlled by the transcription of intergenic sequences have relied on structural functions of the non-coding RNAs produced. These RNAs form part of complexes that modify or remodel chromatin, the combination of DNA and its histone packaging, by changing the degree of packaging or the epigenetic code of chromatin — that is, the heritable information carried in forms other than the DNA sequence. These processes can have opposite effects, as illustrated by the RNA-based hyperactivation of the male X chromosome in *Drosophila*⁴ in comparison with the inactivation of one female X chromosome in mammals⁵, and the silencing of specific mammalian imprinted genes⁶.

But there is another gene-regulatory system in which it is not the intergenic RNA, but rather the process of transcription itself, that triggers changes in the packaging of chromatin. At the genetic locus that encodes the human globin proteins, transcription through regulatory sequences activates certain chromatin domains. Possibly, an increase in DNA accessibility is achieved by chromatin-modifying and chromatin-remodelling factors that ride piggy-back on the RNA polymerase complex⁷.

When compared with the findings described by Martens *et al.*¹, this shows nicely that the passage of processive RNA polymerase through regulatory DNA sequences can also have opposite results. With this in mind, it is intriguing to think that transcriptional interference by intergenic

transcription might also trigger changes in the structure of chromatin. Indeed, previous results of Martens and Winston⁸ showed that repression of the *SER3* gene also depends on components of a chromatin-remodelling complex, one called SWI/SNF. Moreover, other studies in yeast⁹ have shown that transcriptional interference is influenced by the positioning of nucleosomes, the building blocks of chromatin, in the promoter region of the regulated gene.

Taken together, these studies highlight the importance of intergenic transcription in regulating gene activity, even in the relatively densely packed genome of yeast. It seems that RNA polymerases are not only required for the production of particular RNA species, but by travelling along DNA they can also control the occupancy of regulatory sites by transcription factors. Widespread transcription of intergenic sequences has also been described in the human genome. Surprisingly, many of these non-coding transcripts seem to be regulated in a manner that is intimately connected to the

transcription of protein-coding genes¹⁰. So the high proportion of non-coding regions in the genomes of higher organisms is probably not due to the accumulation of nonsense DNA, but rather represents the evolution of ever more complicated gene-regulatory systems.

Sabine Schmitt and Renato Paro are at the Centre for Molecular Biology (ZMBH), University of Heidelberg, Im Neuenheimer Feld 282, 69120 Heidelberg, Germany.
e-mail: paro@sun0.urz.uni-heidelberg.de

1. Martens, J. A., Laprade, L. & Winston, F. *Nature* **429**, 571–574 (2004).
2. Greger, I. H., Aranda, A. & Proudfoot, N. *Proc. Natl Acad. Sci. USA* **97**, 8415–8420 (2000).
3. Corbin, V. & Maniatis, T. *Nature* **337**, 279–282 (1989).
4. Meller, V. H. *et al. Curr. Biol.* **10**, 136–143 (2000).
5. Wutz, A., Rasmussen, T. P. & Jaenisch, R. *Nature Genet.* **30**, 167–174 (2002).
6. Sleutels, F., Tjon, G., Ludwig, T. & Barlow, D. P. *EMBO J.* **22**, 3696–3704 (2003).
7. Gribnau, J., Diderich, K., Pruzina, S., Calzolari, R. & Fraser, P. *Mol. Cell* **5**, 377–386 (2000).
8. Martens, J. A. & Winston, F. *Genes Dev.* **16**, 2231–2236 (2002).
9. Valerius, O., Brendel, C., Düvel, K. & Braus, G. H. J. *Biol. Chem.* **277**, 21440–21445 (2002).
10. Cawley, S. *et al. Cell* **116**, 499–509 (2004).

Theoretical immunology

Parasitic turncoat

Rustom Antia and Jacob Koella

Some parasites evade the immune response of their victim by changing their antigenic coat. Surprisingly, it seems that the trick works best if the new coat isn't completely different from the old one.

Antigenic variation is one of the tricks that pathogens have evolved to escape the immune response of vertebrates. The basic idea is simple. When the specific immune responses directed against the pathogen's antigens have reached a sufficiently high level to clear the infection, the pathogen changes its antigens, rendering the immune responses useless. The benefits of such a strategy are obvious: it extends the duration of infection and the potential for transmission. Why, then, don't more pathogens use antigenic variation? One reason is that most pathogens carry many antigenic determinants (epitopes), which are usually encoded by different genes and expressed on different proteins. Changing only one or a few of these epitopes would not be sufficient to evade the immune response.

Parasites such as *Trypanosoma brucei* solve this problem by covering more than 99% of their cell surface with many copies of a single molecule (the variable surface glycoprotein, or VSG). Expressing a new VSG (from a repertoire of hundreds) allows the trypanosome to evade the antibody response directed against the earlier variants¹. The conventional view of this process is that the different variants are selected to minimize levels of cross-reactivity. A high level of

cross-reactivity — the ability of responses directed against one pathogen variant to clear other variants — would make antigenic variation less effective and would therefore lead to shortened durations of infection.

On page 555 of this issue, Recker *et al.*² turn this conventional view around. Using mathematical models, they show that cross-reactive immune responses, rather than reducing the duration of infection, may actually prolong it. The critical feature of the model, which distinguishes it from earlier ones, is that each variant elicits two types of immune response. One response, which is long-lasting, is directed against a major epitope and is unique to the variant; the other, short-lived response is directed against minor epitopes that are shared by several (but not all) variants. The long-lived, specific immune responses limit the persistence of individual variants by their reaction with the major epitopes, and the short-term, cross-reactive immune responses delay the appearance of variants that express similar combinations of minor epitopes. As a result, the expansion of different antigenic variants is spread out over time, increasing the overall duration of infection.

Recker *et al.* apply their model to antigenic variation during the merozoite stage

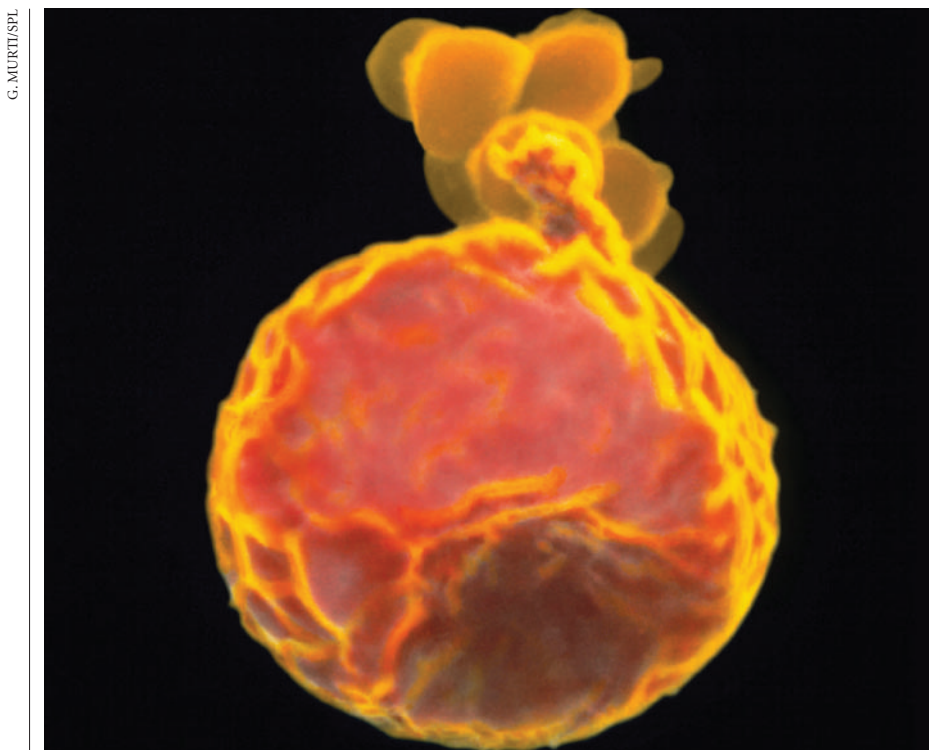


Figure 1 The merozoites of the malaria parasite, *Plasmodium falciparum*, erupt from an infected blood cell, and will carry the infection into other blood cells. Recker *et al.*² suggest that the parasite might have evolved a devious, and counterintuitive, trick to prolong its evasion of antibody responses that are directed against infected blood cells.

of malaria infections, when the parasite (*Plasmodium falciparum*) is replicating inside red blood cells. The parasite then causes an infected blood cell to rupture and release the newly developed merozoites (Fig. 1), which infect other red blood cells. (Synchronous rupture is responsible for the periodic fevers that characterize malaria infections.) Unlike most of the cells in the body, red blood cells do not carry MHC class I proteins on their surface. The merozoites replicating within the cells are therefore not subject to immune surveillance by cytotoxic T lymphocytes. However, a parasite in an infected blood cell is susceptible to being cleared (as are ageing blood cells) as it passes through the spleen. To counter this risk, the parasite inserts PfEMP1 receptors into the surface of the blood cell; this facilitates binding to capillaries and delays passage through the spleen. The PfEMP1 receptors are thus the major target for antibody-mediated immune responses, and have consequently evolved to undergo antigenic variation during the course of the infection.

The model² can explain a number of epidemiological puzzles — for example, that the duration of malaria infection is longer in older than in younger children — as well as accounting for the delayed expansion of different antigenic variants despite the very high antigenic switching rate (which might otherwise be expected to result in nearly synchronous expression of the variants)³.

But Recker *et al.* have not yet compared the predictions of their model with other models of antigenic variation^{4,5}.

Much remains to be done before we understand the dynamics of the merozoite stage of malaria infections. A prime task is to

combine models of antigenic variation with those that allow for the effects of anaemia (caused by rupture of infected blood cells)⁶ and fevers⁷, and to estimate the relevant parameters from experimental data. Different models for the dynamics of antigenic variation must be compared, and we also need a better quantitative understanding of the immune responses elicited by the pathogen — in particular, why certain antigens elicit a greater response (immunodominance) and which factors are responsible for transient responses to some epitopes and for long-lasting immunological memory of others.

Recker and colleagues' model² is an elegant twist on our current ideas about antigenic variation in malaria. But a closer interplay between different models and carefully designed experiments (perhaps using malaria-infected laboratory mice) will be essential in finding the route forward. ■

Rustum Antia is in the Department of Biology, Emory University, Atlanta, Georgia 30322, USA.

e-mail: rantia@emory.edu

Jacob Koella is at the Laboratoire de Parasitologie Evolutive, Université P. & M. Curie, 75252 Paris, France.

e-mail: jkoella@snv.jussieu.fr

1. Barry, J. D. *Parasitology* **92**, 51–65 (1986).
2. Recker, M. *et al. Nature* **429**, 555–558 (2004).
3. Roberts, D. J. *et al. Nature* **357**, 689–692 (1992).
4. Paget-McNicol, S., Gattton, M., Hastings, I. & Saul, A. *Parasitology* **124**, 225–235 (2002).
5. Molineaux, L. *et al. Parasitology* **122**, 379–391 (2001).
6. Gravenor, M. B., McLean, A. R. & Kwiatkowski, D. *Parasitology* **110**, 115–122 (1995).
7. Kwiatkowski, D. & Nowak, M. *Proc. Natl Acad. Sci. USA* **88**, 5111–5113 (1991).

Global change

Hydrocarbon-driven warming

Gerald R. Dickens

A dramatic historical episode of global warming seems to have been driven by the release of huge amounts of hydrocarbons. New evidence for what might have happened comes from the sea floor off Norway.

The outstanding examples of intense global warming and massive greenhouse-gas emissions occurred during a brief episode, known as the 'initial Eocene thermal maximum' (IETM), about 55 million years ago. Superimposed on already warm climates, Earth's surface temperatures soared by 5–10 °C within a geological instant^{1,2}. At the same time, an enormous amount of carbon dioxide, apparently produced through oxidation of hydrocarbons, rapidly entered the global carbon cycle³. Scientists have only reluctantly taken the IETM as an analogue for examining our planet's future, however, because direct evidence for the actual release of hydrocarbons, and the driving mechanism, has remained elusive.

This will change if the observations and ideas of Svensen and colleagues⁴ (page 542 of this issue) prove to be correct.

There is seemingly incontrovertible evidence of the injection of huge amounts of organically derived CO₂ during the IETM. Numerous isotope records, constructed using data from primary carbonate or organic matter, display an extraordinary drop in the ratio of ¹³C to ¹²C at that time, clearly signalling a perturbation of the entire carbon pool on Earth's surface⁵. Given the coeval dissolution of deep marine carbonate^{5,6}, and the short duration of the isotope excursion (< 20,000 years^{1,5–7}), at least 1,500 gigatonnes (Gt) of carbon as CO₂ must have suddenly been injected into the ocean or atmosphere

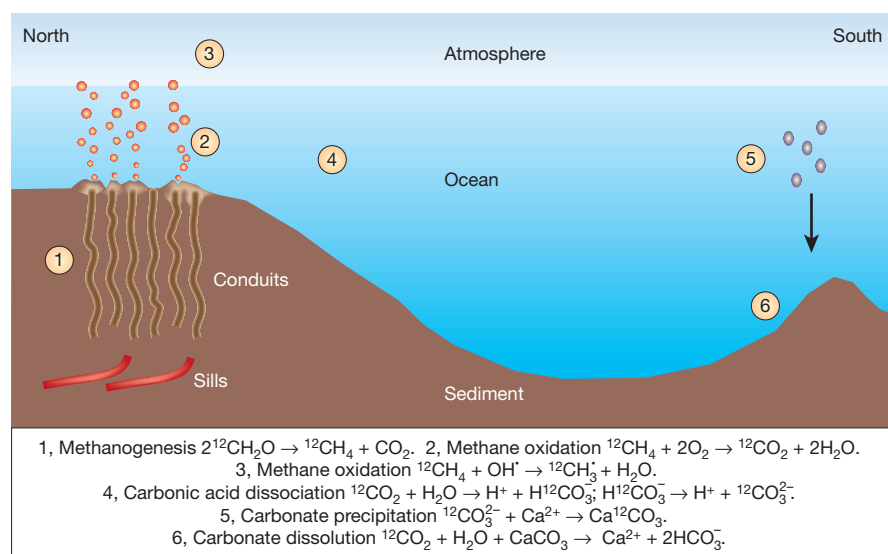


Figure 1 Methane release during the initial Eocene thermal maximum (IETM), around 55 million years ago. This overall scheme incorporates both sediment-core data^{5,8} and the new seismic observations of Svensen *et al.*⁴. Sediment with abundant ^{13}C -depleted organic carbon ($^{12}\text{CH}_2\text{O}$) was penetrated by hot sills, resulting in the release of massive amounts of newly formed or previously generated ^{13}C -depleted methane ($^{12}\text{CH}_4$). Conduits channelled the gas–fluid mixture to the sea floor, and the CH_4 was then oxidized to ^{13}C -depleted carbon dioxide ($^{12}\text{CO}_2$) in the ocean and atmosphere, with consequent extreme warming. Such oxidation resulted in finely layered strata devoid of carbonate in the North Atlantic, and a global negative carbon-isotope anomaly⁹: these phenomena are seen in sediment cores at the onset of the IETM. The numbered sequence shows the chain of generic reactions involved.

from a ^{13}C -deficient source³. The only satisfactory explanation for this is the release and oxidation of carbon from a large reservoir of organic material (the biological generation of organic compounds preferentially incorporates ^{12}C).

In several sediment sections from the north and central Atlantic Ocean, the IETM is marked by finely layered deposits devoid of carbonate^{5,8}. Combined with the isotope information, that suggests the addition of substantial amounts of methane to this ocean basin, and its oxidation (Fig. 1), because methane oxidation consumes dissolved oxygen (which precludes fauna from turning the sediment over) and produces CO_2 (which dissolves carbonate)⁹. The dual discoveries of the apparent existence of warm, deep waters at the IETM¹, and vast amounts of methane-containing gas hydrates — ice-like crystals of gas and water — along continental slopes¹⁰, thus led to the following proposal^{3,5,6,9}: like those of today, gas hydrates in marine sediments included large amounts of ^{13}C -depleted biogenic methane; some change in conditions at the start of the IETM caused relatively warm water to sink in the oceans, and that warming resulted in the dissociation of the gas hydrate; the free methane thus released escaped from the sea floor, and was then oxidized to CO_2 in the ocean or atmosphere.

Such a massive release of methane should have left physical traces, especially in North Atlantic sediments where the chemical evidence for such an event is strongest⁹. However, other than sediment slumping along the North American margin⁶, which might be explained by processes other than gas release, no such evidence has been forthcoming. There are also problems with identifying a plausible environmental trigger for hydrocarbon release from marine sediments^{3,11}. Alternative, but similarly unsatisfactory, explanations for the carbon input have included the impact of a volatile-rich comet¹¹

or burning of expansive peat deposits¹².

Svensen *et al.*⁴ have examined data from an immense seismic array that shows features lying beneath the sea floor in two large sedimentary basins west of Norway. A truly remarkable find is their identification of several hundred kilometre-scale ‘fluid-escape conduits’ that terminate in marine strata deposited around 55 million years ago. A few such structures had been previously identified in the vicinity and were interpreted as submarine igneous volcanoes¹³. But these conduits, which are capped by craters and

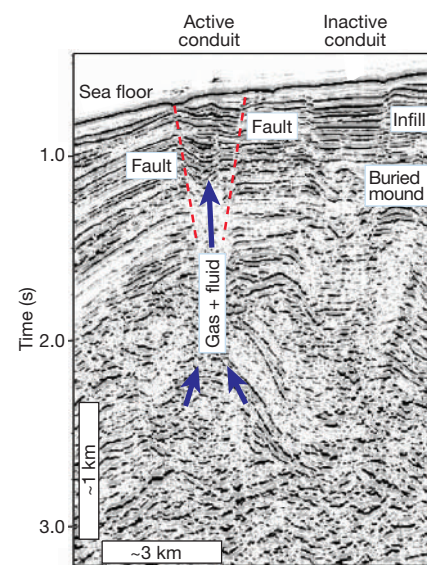


Figure 2 Seismic profile of active and inactive ‘fluid-escape conduits’ on the upper continental slope southwest of the modern Niger River Delta. These features result from the migration of water, gas and sediment from depth, and resemble the 55-million-year-old structures identified by Svensen *et al.*⁴ in the northeast Atlantic (see Fig. 2 on page 543). Note collapse of sediment along faults above both conduits, and sediment burial of the mound at the inactive conduit. The vertical axis is the two-way travel time of sound waves in seconds. (Reproduced from ref. 14.)

mounds, more closely resemble structures seen in modern sedimentary basins, where water and gas migrate from depth and spawn mud volcanoes and hydrocarbon seeps on the sea floor (Fig. 2)^{14,15}. The ancient structures seem to be connected to volcanic sills^{4,13}, which delivered hot magma into organic-rich sediments throughout the North Atlantic around 55 million years ago. So the authors⁴ argue that, during the IETM, the sills produced massive amounts of thermogenic methane and high-pressure fluids, which vented into the ocean through the conduits and perturbed the global carbon cycle (Fig. 1).

All this provides tantalizing support for massive hydrocarbon discharge from the sea floor during the IETM. The composition of the expelled fluids, and the timing of their release, need to be better defined, however. And a convincing link between the conduits and environmental changes at the IETM will require evidence that significant quantities of ^{13}C -depleted carbon were actually produced in and released from sedimentary rather than volcanic¹³ features, and that this occurred during the 20,000-year onset of the carbon isotope anomaly.

The simultaneous generation and release of methane by volcanic sills at the IETM solves the triggering problem. But it raises an equally contentious issue. Thermogenic methane resulting from the heating of sedimentary organic carbon contains a greater proportion of ^{13}C than does biogenic methane. So to explain the isotope anomaly, a greater amount of carbon than previously thought — more than 3,000 Gt — must have escaped^{3,4}. The entire world’s conventional deposits of oil and gas today amount to about 5,000 Gt of carbon¹⁰. The mechanism proposed by Svensen *et al.*⁴ therefore requires the instantaneous generation and release of a truly stupendous amount of hydrocarbons. An alternative possibility is that the sills released biogenic methane that

had already accumulated in North Atlantic strata and then, through the consequent environmental changes, carbon from other sources such as methane from widely dispersed gas hydrates.

A better understanding of the relationship between the sills, conduits and carbon-cycle perturbation at the IETM will require more work. But if that relationship is one of cause and effect, the significance of the IETM escalates dramatically. In the hydrate-dissociation scenario^{3,6}, deep-ocean warming drove the massive release of carbon, making events at the IETM an intriguing but imperfect analogue of current fossil-fuel emissions. The volcanic triggering of methane release from the sea floor, whether that methane was biogenic or thermogenic, instead implies that sudden hydrocarbon input caused extreme warming, a view consistent with analyses² of temperatures at the IETM. Given the comparable estimates for carbon release at the IETM (1,500 to 3,000 Gt)^{3,4}, and anthropogenic release of

carbon into the atmosphere over the coming centuries (3,000–4,000 Gt), environmental change during the IETM should become the subject of general investigation. ■

Gerald R. Dickens is in the Department of Earth Sciences, Rice University, Houston, Texas 77251, USA. e-mail: jerry@rice.edu

1. Kennett, J. P. & Stott, L. D. *Nature* **353**, 225–229 (1991).
2. Zachos, J. C. *et al.* *Science* **302**, 1551–1554 (2003).
3. Dickens, G. R., O'Neil, J. R., Rea, D. K. & Owen, R. M. *Paleoceanography* **10**, 965–971 (1995).
4. Svensen, H. *et al.* *Nature* **429**, 542–545 (2004).
5. Bralower, T. J. *et al.* *Geology* **25**, 963–966 (1997).
6. Katz, M. E., Pak, D. K., Dickens, G. R. & Miller, K. G. *Science* **286**, 1531–1533 (1999).
7. Farley, K. A. & Eltgroth, S. F. *Earth Planet. Sci. Lett.* **208**, 135–148 (2003).
8. Shipboard Scientific Party. ODP Prelim. Rep. 107 www-odp.tamu.edu/publications/prelim/207_prel/207PREL.PDF (2003).
9. Dickens, G. R. *Bull. Soc. Geol. Fr.* **171**, 37–49 (2000).
10. Kvenvolden, K. A. *Chem. Geol.* **71**, 41–51 (1988).
11. Katz, D. V. *et al.* *Earth Planet. Sci. Lett.* **211**, 13–26 (2003).
12. Kurtz, A. C., Kump, L. R., Arthur, M. A., Zachos, J. C. & Paytan, A. *Paleoceanography* **18**, 1090 (2003).
13. Davies, R., Bell, B. R., Cartwright, J. A. & Shoukors, S. *Geology* **30**, 223–226 (2002).
14. Graue, K. *Mar. Petrol. Geol.* **17**, 959–974 (2000).
15. Dimitrov, L. I. *Earth Sci. Rev.* **59**, 49–76 (2002).

Neurobiology

A matter of balance

Martyn Goulding

The types of chemical signal that a neuron synthesizes and communicates with were thought to be genetically encoded and largely invariable. It seems, though, that if a neuron's activity changes, so too do its signals.

Nerve cells use chemical signals known as neurotransmitters to communicate with each other. These molecules come in many different flavours, and the combination of flavours used by any given neuron represents a key property that determines not only its function within a circuit, but also the circuit's overall output. How exactly neurons regulate the profile of neurotransmitters that they express — their neurotransmitter 'phenotype' — is poorly understood, although it is known that most neurons synthesize a highly restricted repertoire of neurotransmitters, and that the regulatory events governing this repertoire occur in the embryo, soon after a neuron is born and begins to take on a particular identity. Numerous studies^{1–4} have also led to the view that a neuron's transmitter phenotype is tied closely to the genetic programme that controls its developmental fate, such that a neuron's identity and its neurotransmitter profile are inextricably interwoven.

On page 523 of this issue⁵, however, Borodinsky and colleagues challenge this view. These authors provide evidence that, in the spinal cord of developing frogs, profiles of neurotransmitter expression can change in response to differing degrees of neuronal activity. Although it was well known that

neurons can alter the levels of expression of particular neurotransmitters after changes in circuit activity, what is interesting about the new study is that it shows that embryonic spinal-cord neurons can also alter the types of neurotransmitter that they produce — and that they do this independently of changes in cell identity. Moreover, these changes in neurotransmitter phenotype, which seem to be encoded by the patterns of Ca²⁺ spikes — a measure of activity — that the neurons produce, occur in a system that was thought to be genetically 'hardwired'.

Previously, patterns of Ca²⁺ spikes were shown to modulate the expression of neurotransmitters *in vitro*⁶. Borodinsky *et al.*⁵ have now taken an elegant approach to assessing the role of Ca²⁺-dependent activity in neurotransmitter expression *in vivo*. First, they noted that different populations of embryonic neurons exhibit distinctive patterns of Ca²⁺ spikes. Then they studied the effects of experimentally manipulating this activity, by engineering the developing spinal cord to overexpress one of two types of ion channel: a potassium channel that hyperpolarizes neurons and so reduces Ca²⁺ spike activity, or a sodium channel that increases the frequency of spike activity. By injecting messenger RNA transcripts encoding one of these channels at the two-cell stage of embryonic



100 YEARS AGO

In the course of an interview reported in the *Westminster Gazette* of Friday last, Lord Kelvin is reported to have expressed himself as being decidedly of the opinion that the source of energy of the heat emitted by radium is not in the element itself. He remarked:—"It seems to me absolutely certain that if emission of heat at the rate of 90 calories per gram per hour found by Curie at ordinary temperature, or even at the lower rate of 38 found by Dewar and Curie from a specimen of radium at the temperature of liquid oxygen, can go on month after month, energy must somehow be supplied from without."

ALSO

A Reuter message from Wellington, New Zealand, reports that the King has sent the following telegram to Captain Scott, leader of the National Antarctic Expedition:—"I have read with interest your report, which Sir Clements Markham sent me. I congratulate you and your gallant crew on your splendid achievements, and wish the *Discovery* a safe journey home. I hope to see you on your return to England."

From *Nature* 2 June 1904.

50 YEARS AGO

It is no bad thing that a broadsheet entitled "Graduate wives" ... has provoked considerable discussion regarding the value of university education for women, and perhaps more particularly when that education has only been possible because of the help received from public funds in some form or another. That latter point was not specifically covered by the inquiry, although it is noted that many thousands of pounds of public money are spent on the education of the four thousand women graduates who leave the universities of Great Britain every year... the fact that the majority of those covered by this inquiry made no direct use of their academic qualifications after marriage does not imply that the public money expended on their university education has been wasted, though it may well induce some re-examination of the system of university awards and a fuller consideration of the whole purpose of university education. As the broadsheet points out, the indirect contribution which a trained mind and cultured outlook can make to family life and to the life of the nation generally is very great indeed.

From *Nature* 5 June 1954.

development, the authors could manipulate spike activity in either one or both halves of the spinal cord when it formed later on in development.

Intriguingly, they find that a decrease in Ca^{2+} spike activity is correlated with an increase in the number of neurons that produce marker proteins indicative of the expression of glutamate and acetylcholine — two types of excitatory neurotransmitter. In contrast, increased spike activity results in fewer cells expressing these neurotransmitters. Similarly, in their initial experiments Borodinsky *et al.* found that, at early stages of development, motor neurons — which produce acetylcholine — normally generate a low frequency of spike activity. Meanwhile, ventral interneurons, another type of nerve cell, show a high rate of activity and produce little acetylcholine. But when sodium channels were overexpressed in nerve cells that include inhibitory commissural neurons, more ventral interneurons expressed ChAT, the enzyme that catalyses acetylcholine production. This presumably reflects a decrease in the frequency of Ca^{2+} spikes in these ventral interneurons, because of increased inhibition by commissural neurons that are generating more spikes than usual.

The authors also observe reciprocal changes in the expression of the inhibitory neurotransmitters γ -aminobutyric acid (GABA) and glycine, such that fewer neurons express these neurotransmitters when Ca^{2+} spike activity is low, and more neurons do so when spike activity is high. Remarkably, increases and decreases in activity also result in excitatory and inhibitory neurotransmitters being expressed at the same time in some neurons. Somewhat surprisingly, none of these changes affect the expression of certain markers of neuronal identity, suggesting that a neuron's identity and its neurotransmitter phenotype are not strictly linked. Together, these findings suggest that a homeostatic mechanism, one that attempts to maintain the status quo, regulates the overall level of neuronal activity in the embryonic spinal cord.

An indication that these activity-dependent changes might be important in normal development comes from the authors' observation that there is an essential period during which Ca^{2+} spike activity is required for altered neurotransmitter expression — a phenomenon seen in other systems in which neural activity regulates circuit formation. Nevertheless, the extent to which activity controls neurotransmitter expression during normal development remains to be seen. Although Borodinsky and colleagues' results show that perturbations in Ca^{2+} spike activity can clearly, in an experimental set-up, override the genetic programme that determines neurotransmitter phenotype, the authors of another recent study³ argue that a genetically hardwired programme operates

in the mouse spinal cord to direct the formation of glutamate-producing as opposed to GABA-producing neurons.

Borodinsky *et al.* further show that activity-related changes in neurotransmitter expression also occur in isolated neurons *in vitro*, hinting that a cell-autonomous mechanism — a mechanism that depends solely on the neuron in question, and not on external factors — regulates neurotransmitter output. So is the activity-dependent regulation of neurotransmitter expression a property of individual neurons, or of the circuit to which they contribute, or both?

Interestingly, the neurons that exhibit activity-induced changes in frogs are constituents of the primary nervous system that controls larval swimming and escape reflexes. As these circuits develop very early, and are 'wired-up' during the essential period of Ca^{2+} spike activity, one could imagine a mechanism that operates to calibrate the overall activity of neurons in this circuit, and thus ensure that larval movements are properly coordinated. In embryos that do not use such motor behaviours, this aspect of regulation might be missing. Furthermore, as GABA and glycine are initially excitatory in the spinal cords of birds and mammals, changes in the expression of these neurotransmitters would not necessarily alter the balance between excitation and inhibition.

Behavioural genetics

All in the family

Allen J. Moore

Mothers and offspring may have different ideas about how much maternal care should be provided. How is the behaviour of both parties genetically influenced, and how is this evolutionary conflict resolved?

Family conflicts are of professional interest not only to agony aunts but to evolutionary biologists as well. A case to exercise the latter has just appeared in *Proceedings of the Royal Society*¹, where James Curley and colleagues describe their manipulations of a mouse gene, *Peg3*. This gene is of particular interest because it is known to influence maternal behaviour, but is also inherited in an unusual fashion.

Conflicts between mother, father and offspring take several forms. Newborns in a species that provides parental care often demand more than a mother is willing to give. The interest of the mother is to ensure that the needs of her offspring are met while retaining the ability to take care of other or future offspring; the interest of an offspring is to get as many resources as possible for itself, regardless of the interests of its siblings or mother. Fathers' interests may coincide with those of the offspring, particularly if there is uncertainty that he will father other

These findings⁵ raise several additional questions. First, how widespread is this mechanism of activity-dependent neurotransmitter regulation? Second, how do activity-dependent regulatory pathways interface with the genetic programmes that control neurotransmitter phenotype? Third, what is the underlying molecular mechanism that translates Ca^{2+} spike activity into changes in neurotransmitter expression? Last, are alterations in neurotransmitter expression matched by alterations in the expression of their cognate receptors on downstream neurons? Clearly there is a lot more to know about the role of activity and Ca^{2+} signalling in neuronal differentiation. Nonetheless, Borodinsky and colleagues' study indicates that concentrating solely on the molecular and genetic mechanisms that control neurotransmitter expression will cause us to overlook some of the essential pathways that configure neural circuits. ■

Martyn Goulding is in the Molecular Neurobiology Laboratory, Salk Institute for Biological Studies, La Jolla, California 92037, USA.
e-mail: goulding@salk.edu

1. Goridis, C. & Rohrer, H. *Nature Rev. Neurosci.* **3**, 531–541 (2002).
2. Hendricks, T. J. *et al. Neuron* **37**, 233–247 (2003).
3. Cheng, L. *et al. Nature Neurosci.* **7**, 510–517 (2004).
4. Pattyn, A. *et al. Genes Dev.* **17**, 729–737 (2003).
5. Borodinsky, L. *et al. Nature* **429**, 523–530 (2004).
6. Watt, S. D., Gu, X., Smith, R. D. & Spitzer, N. C. *Mol. Cell Neurosci.* **16**, 376–387 (2000).

or future offspring in this family and the mother provides all the care. Conflicting selection pressures can therefore arise. If the behaviour is genetically influenced, that raises questions about how such family conflicts are resolved over evolutionary time².

Curley and colleagues¹ have applied modern techniques to investigate the genetics of such conflicts in mice. The *Peg3* gene they studied is paternally expressed: that is, the mother's copy of the gene is turned off in the offspring. This phenomenon, known as 'imprinting', is predicted to have evolved in genes whenever there are asymmetries in the evolutionary interests of parents in the behaviour directed towards offspring³. But whose behaviour is affected? Is the action of *Peg3* expressed in mothers or in offspring?

Because individuals only express the *Peg3* gene they inherit from their father, it is possible to generate mouse families in which the mother expresses the functional *Peg3* gene product but the offspring don't, and vice

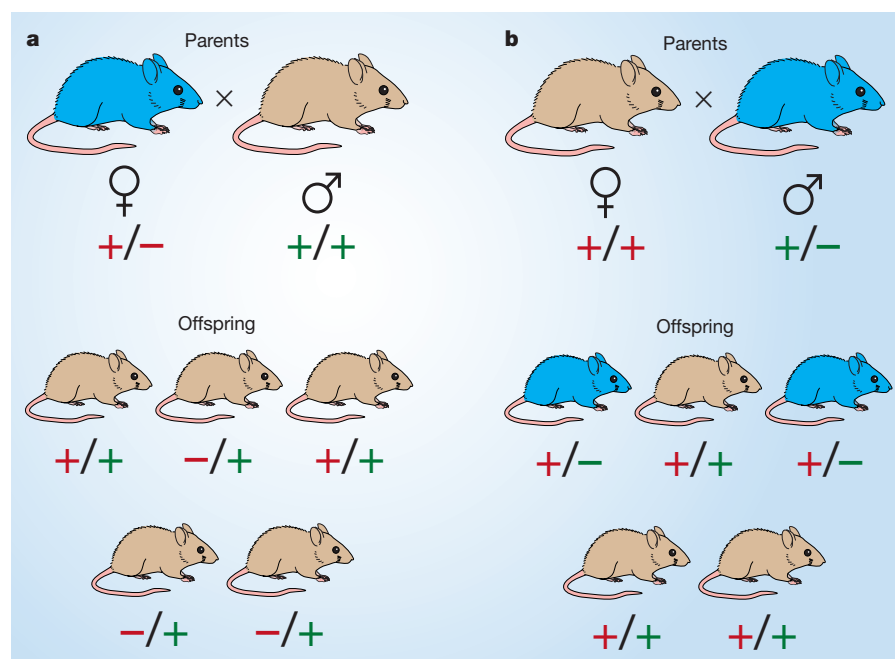


Figure 1 Expressing *Peg3* or a reporter gene. **a**, In this cross, the mother expresses a reporter gene (β -galactosidase, which produces blue pigment) and not *Peg3* because she has inherited the mutated gene ($-$) from her father. She also has a silent copy of the wild-type *Peg3* gene ($+$) from her mother. When crossed to a fully wild-type male, her offspring will all express wild-type *Peg3* because they inherit the wild-type gene copy from their father, regardless of which copy they inherit from their mother. In this cross, a mutant mother is raising normal babies. **b**, A normal mother is mated to a *Peg3* mutant male that expresses the reporter gene he inherited from his own father. The babies from this cross will inherit a silent wild-type gene copy from their mother. But some will have their father's mutant copy and will thus express the reporter gene, while others will have their father's wild-type gene and will be normal. Here, a normal mother is raising a mixture of normal babies and babies that don't express *Peg3*. Other crosses were made that resulted in litters with all mutant offspring or offspring with two wild-type copies of the gene.

versa (Fig. 1). To ensure that the gene is silenced, and to recognize which individuals are expressing *Peg3* and which are not, Curley *et al.* used mice with a 'targeted mutation', where the function of *Peg3* is disrupted by insertion of a 'reporter gene' that turns cells blue if the new gene is expressed instead of *Peg3*. The mutation was created in an otherwise genetically identical inbred strain of mice, so that the only difference was the presence or absence of *Peg3* expression. This is a powerful experimental approach, as mothers can have mixed litters of offspring that express or lack *Peg3* (Fig. 1).

Using these lines, the authors looked for the gene's effects on maternal behaviour by comparing mutant mothers (those that had the reporter gene instead of *Peg3*) and wild-type mothers (those with functional *Peg3*) rearing all wild-type offspring. They also compared wild-type and mutant offspring reared by wild-type mothers to examine how *Peg3* influenced offspring traits.

The results were striking on both counts. Although they had the same number of pups, mutant mothers ate less and gained less weight in their pregnancy, had less fat reserves after giving birth, and produced less milk. Mutant pups were less competent at suckling, and performed less well even after

weaning. They also lost out in sibling competition, as pups born into families with a mixture of wild-type and mutant offspring were smaller than counterparts born into exclusively wild-type or mutant families. The point here is that asymmetry among siblings resulted in greater competition compared with families of all equals.

Finally, newborn mice depend on their mother to keep their body temperature constant and high. Mothers lacking *Peg3* were poor nest-builders, and failed to retrieve offspring that wandered from the nest; *Peg3*-deficient offspring were delayed in developing self-thermoregulation. In the end, absence of *Peg3* expression led to smaller pups that had delayed reproduction and were less likely to survive. *Peg3* therefore affects the behaviour of both mother and offspring.

Genes such as *Peg3* that influence the fitness of both parents and offspring require us to consider simultaneous evolution in two generations, rather than the more typical single generation at a time. However, such shared genetic influences on mothers and offspring are consistent with previous theoretical^{4,5} and empirical work⁶⁻⁹ showing that genetic coadaptation of parent and offspring behaviour is to be expected and can evolve. A positive coadaptation as shown by Curley

et al., where offspring that are able to get the most from their mothers grow up to be good providers, is also seen in parental care and offspring begging in a bird⁶ and an insect⁹. So selection acts through both the mothers and the offspring⁹. However, in the case of *Peg3*, only the father's gene is selected, so mothers may be working harder than they might if both copies of the gene were expressed.

Demonstrating the effect of a gene using targeted mutagenesis (the focus of molecular behaviour genetics) is not the same as demonstrating genetic influences on natural variation in behaviour (the focus of evolutionary behaviour genetics). In addition, parenting and infant behaviour are likely to be influenced by many genes rather than by a single gene. Evidence for the effects of natural variation in *Peg3* on parent-offspring interactions in wild populations must await further study, but other genetic investigations of parental performance in different strains of mice also implicate *Peg3* as one of many influences on parenting and infant survival¹⁰. Thus, the effect is not solely due to a complete disruption of the gene: normal levels of variation at this genetic locus seem to be important, at least in differences among laboratory strains.

Is the behaviour of mother and offspring evolving because of conflict or as a coadaptation? The studies of *Peg3* in mice provide direct evidence that parent and offspring behaviour can have a common genetic basis, and support the idea that both conflict and coadaptation between parents and offspring can be accommodated. Or, as today's parents learned from the Rolling Stones, you may not always get what you want, but you might just find that you get what you need.

Allen J. Moore is in the School of Biological Sciences, University of Manchester, Manchester M13 9PT, UK. e-mail: allen.j.moore@man.ac.uk

- Curley, J. P., Barton, S., Surani, A. & Keverne, E. B. *Proc. R. Soc. Lond. B* doi:10.1098/rspb.2004.2725 (2004).
- Parker, G. A., Royle, N. J. & Hartley, I. R. *Phil. Trans. R. Soc. Lond. B* **357**, 295–307 (2002).
- Haig, D. *Proc. R. Soc. Lond. B* **264**, 1657–1662 (1997).
- Wolf, J. B. & Brodie, E. D. III *Evolution* **52**, 299–308 (1998).
- Wolf, J. B. *Evolution* **54**, 1882–1898 (2000).
- Kölliker, M. & Richner, H. *Anim. Behav.* **62**, 395–407 (2001).
- Agrawal, A. F., Brodie, E. D. III & Brown, J. *Science* **292**, 1710–1712 (2001).
- Hager, R. & Johnstone, R. A. *Nature* **421**, 533–535 (2003).
- Lock, J. E., Smiseth, P. T. & Moore, A. J. *Am. Nat.* (in the press).
- Peripato, A. C. & Cheverud, J. M. *Am. Nat.* **160**, S173–S175 (2002).

Correction

In Julian Kroklik's News and Views article, "Dust-filled doughnuts in space" (*Nature* **429**, 29–30; 2004), it was incorrectly stated that the observations described were the first to be made of an extragalactic source using infrared interferometry. In fact, the first such observations were published by M. Swain *et al.* (*Astrophys. J.* **596**, L163–L166; 2003), using the Keck interferometer to resolve the core of NGC 4151 at 2.2 micrometres.

Photonics

A new source of terahertz radiation?

Appl. Phys. Lett. **84**, 4810–4812 (2004)

Imaging using terahertz radiation, which has wavelengths intermediate between those of infrared and microwaves, could provide an alternative to the conventional techniques that use X-rays or ultrasound. Terahertz (THz) imaging could be cheap, non-hazardous and chemically specific. But it requires strong sources of THz radiation. Although semiconductor materials that emit at visible and near-infrared wavelengths are well known, it is harder to find semiconductors with bandgaps small enough to produce THz radiation. (The larger the bandgap, the shorter the emission wavelength.)

Now Ricardo Ascáubi and colleagues report that indium nitride (InN) is a promising THz source. This material is still rather patchily characterized, and previous measurements of the bandgap have given conflicting results. Ascáubi *et al.* show that thin films of InN grown on sapphire and excited with infrared laser pulses emit radiation with a peak at around 1 THz, consistent with the idea that InN has a very small bandgap. Already the researchers can achieve emission intensities comparable to that of indium arsenide, one of the strongest-known THz emitters. They anticipate that InN should perform even better if the crystallinity of the thin films can be improved.

Philip Ball

Medicine

Diabetic genetics

Science **304**, 1325–1328 (2004)

Most forms of diabetes are caused by a suite of different mutations rather than a single identifiable one. But some rare forms are the result of a single gene defect, allowing researchers a glimpse into the mechanisms that might be involved in the more common versions.

Stella George and colleagues screened DNA from patients with severe insulin resistance, in which the body is impervious to the sugar-regulating effects of insulin. One person, who suffered from hereditary late-onset diabetes, had a mutation in a gene, *AKT2*, which encodes a so-called protein kinase involved in mediating insulin action. The patient's mother, grandmother and maternal uncle all had the mutation, and diabetes, whereas some 1,500 control subjects did not.

The discovery implicates *AKT2* in regulating blood sugar, and more generally shows the potential importance of individual components of that pathway. Its broader relevance to more common

forms of diabetes, however, would require identification of more diabetics with the same mutation.

Michael Hopkin

Chemistry

Solid prospects

Angew. Chem. Int. Edn **43**, 2955–2958 (2004)

Many large-scale chemical processes are catalysed by sulphuric acid. But about 14 million tonnes of acid are wasted each year because the recycling process is inefficient. A solid acid would be easier to separate from a mixture of liquids and, if it could be produced cost-effectively for industrial use, would be cheaper and 'greener'.



With this in mind, Michikazu Hara *et al.* set to work, and they have synthesized a carbon-based solid acid that is twice as effective as sulphuric acid in catalysing a standard hydrolysis reaction. In other common reactions the solid acid was slightly less effective, but still had comparable activity. The catalyst's activity and composition remained unchanged after four reaction cycles.

The solid acid is formed in an inexpensive process in which naphthalene is heated in sulphuric acid at 300 °C. The resulting black powder is insoluble in common reaction solvents, and contains about five times more proton-delivering groups — and so potential catalytic power — than an existing commercial solid acid used in laboratory work. The authors speculate that their acid may also be useful as a proton carrier in certain types of fuel cell.

Mark Peplow

Gene therapy

RNA versus brain cancer

Clin. Cancer Res. **10**, 3667–3677 (2004)

Small RNAs are important modulators of gene expression. These molecules bind to complementary regions of protein-encoding messenger RNAs, and either promote their degradation or inhibit their translation into

protein. Thus, small RNAs offer the promise of being highly specific drugs. Yun Zhang and colleagues have now demonstrated the potential of these molecules for treating brain cancer.

Some 90% of primary and metastatic brain cancers rely for their growth on the epidermal growth factor receptor (EGFR), so molecules that block this protein are attractive anticancer drugs. Generally, such drugs would be administered intravenously, but the blood–brain barrier poses an obstacle to this; blood vessels surrounding brain tumours have restricted permeability to cells and molecules.

Zhang and colleagues overcame this impediment by placing two antibodies on the surface of a drug-delivery vehicle. One antibody allowed it to pass through the walls of the cancer vasculature, and the other promoted its entry into cancer cells. Once inside the cells, DNA containing the gene for a small RNA that inhibits EGFR was released, and the small RNA was made. This suppressed 95% of EGFR activity and significantly increased the survival time of mice with implanted human brain tumours.

Angela K. Eggleston

Optical tweezers

A light twist

Phys. Rev. Lett. **92**, 190801 (2004)

Light can be used like a wrench to twist tiny objects. In an optical trap, an intense light field exerts linear forces on an object to fix it in space. If the light is polarized, the force can be given a twist. Arthur La Porta and Michelle D. Wang have used this principle to apply a precisely defined torque to a speck of quartz.

The trapping force in an optical trap is created by electrical polarization of the object in an electromagnetic field. Because quartz is doubly refracting, this polarization is not the same in all directions, and in a polarized trapping beam the material will tend to align its optical axis with the polarization of the beam, creating a torque. Alexis I. Bishop *et al.* used this idea previously to rotate microscopic glass cylinders optically (*Phys. Rev. A* **68**, 033802; 2003).

La Porta and Wang show that they can rotate a 1- μm quartz particle in this way, and that, by measuring the angular momentum of the transmitted beam, they can calculate the torque transmitted to the particle. By adding a feedback mechanism, they adjust the polarization angle of the beam driving the twist to maintain a constant torque on the particle. Such a scheme could be used to measure and to modulate the torque exerted by biological rotary motors such as ATP synthase, which are known to generate torques within the dynamic range of this apparatus.

Philip Ball

Where leatherback turtles meet fisheries

Conservation efforts should focus on hot spots frequented by these ancient reptiles.

The dramatic worldwide decline in populations of the leatherback turtle (*Dermochelys coriacea*)¹ is largely due to the high mortality associated with their interaction with fisheries², so a reduction of this overlap is critical to their survival. The discovery of narrow migration corridors used by the leatherbacks in the Pacific Ocean³ raised the possibility of protecting the turtles by restricting fishing in these key areas. Here we use satellite tracking to show that there is no equivalent of these corridors in the North Atlantic Ocean, because the turtles disperse actively over the whole area. But we are able to identify a few 'hot spots' where leatherbacks meet fisheries and where conservation efforts should be focused.

The Atlantic beaches of French Guiana and Suriname (5° 5' N, 54° W) host the last major nesting sites for leatherback turtles¹ (Fig. 1). But because the oceanic migration routes of these animals are largely unknown, the development of spatially focused conservation strategies is hindered. We deployed Argos satellite transmitters on leatherback turtles nesting in French Guiana ($n=29$ turtles) and Suriname ($n=3$; turtles H, J and K) and monitored their movements during ($n=20$ individuals) and after ($n=12$) nesting seasons from 1999 to 2002.

Migration trajectories (Fig. 2) indicate that, unlike in the Pacific³, leatherbacks disperse widely throughout the North Atlantic Ocean. They follow at least two

main patterns, which depend on the year: some disperse north, broadly towards the Gulf Stream area, and others disperse to the east and remain in tropical waters.

Turtles migrating north (turtles E, F, G and H in Fig. 2) swam with strikingly constant individual headings (range: $7.1 \pm 4.2^\circ$ to $348.3 \pm 4.0^\circ$). As they did so, they cut across the whole subtropical gyre, generally moving at an angle across the main current. After crossing the Gulf Stream, two individuals (E and G) veered to the east into the plankton-rich transition zone between the subtropical and the subpolar gyres. Their behaviour then changed markedly and became closely correlated with the local oceanographic conditions as their trajectories curved along the oceanic fronts, which are perfectly recorded in the contemporaneous satellite altimeter data. (For movie, see supplementary information.)

The turtles that were following oceanic fronts travelled more slowly than when they were moving straight towards the north (for example, turtle G: $0.57 \pm 0.03 \text{ m s}^{-1}$, $n=183$, where n is the number of speeds calculated between locations, as opposed to $0.68 \pm 0.04 \text{ m s}^{-1}$, $n=135$; t -test, $t=2.239$ and $P=0.026$). This suggests that turtles slow down to forage along productive fronts where their main prey, gelatinous plankton, is concentrated⁴. This is also where pelagic, longline fishing boats aggregate^{5,6} and take a number of sea-turtles as bycatch⁷.

Other individuals (B, C, D, I, J and L) swam a short northward leg up to about 10° N and then headed generally eastwards (range: $52.5 \pm 5.4^\circ$ to $101.9 \pm 14.4^\circ$), mostly swimming against the westward flowing North Equatorial Current. Two of them (B and C) were tracked far to the east. Turtle B reached the Cape Verde Islands before transmission stopped. Turtle C followed a



Figure 1 A leatherback turtle leaving the beach of Awala-Yalimapo in French Guiana. Atlantic beaches around this area are considered to be the last major nesting sites for the species.

southern route, crossing the eastward flowing North Equatorial Countercurrent (roughly between 10° N and 3° N) before reaching the westward flowing South Equatorial Current, where she reversed her course, heading west with the current when she reached about 2° N . This turtle probably fed on the productive frontal areas associated with the equatorial current systems. Large fishing fleets operate here, targeting tropical tuna species (*Katsuwonus pelamis*, *Thunnus obesus* and *Thunnus albacares*)⁸.

During nesting, tracking data indicate that turtles remain within 20 km of the front of the Maroni river estuary — an area that is heavily exploited by French Guiana and Suriname fishing fleets catching the brown shrimp *Penaeus subtilis* and the red snapper *Lutjanus purpureus*⁹. Unsurprisingly, frequent accidental sea-turtle catches are reported for these fleets¹⁰.

The few hot spots where turtles are likely to encounter coastal or pelagic fishing fleets are very different and are widely scattered across the Atlantic basin. This finding highlights the pressing need to develop locally adapted, but basin-wide and internationally coordinated, conservation strategies for preserving the last large population of leatherback turtles.

Sandra Ferraroli*†‡§, Jean-Yves Georges*, Philippe Gasparil, Yvon Le Maho*†

*Centre d'Ecologie et Physiologie Énergétiques, Centre National de la Recherche Scientifique, CEPE-CNRS UPR 9010, 67087 Strasbourg, France
e-mail: jean-yves.georges@c-strasbourg.fr

†Université Louis Pasteur, 67070 Strasbourg, France

‡Direction Régionale de l'Environnement, Direction Guyane, 97300 Cayenne, Guyane Française, France
§Present address: WWF-France, 97300 Cayenne, Guyane Française, France

||Direction Océanographie Spatiale, Collecte Localisation Satellites, CLS, 31526 Ramonville, France

1. Spotila, J. R. et al. *Chelon. Conserv. Biol.* **2**, 209–222 (1996).

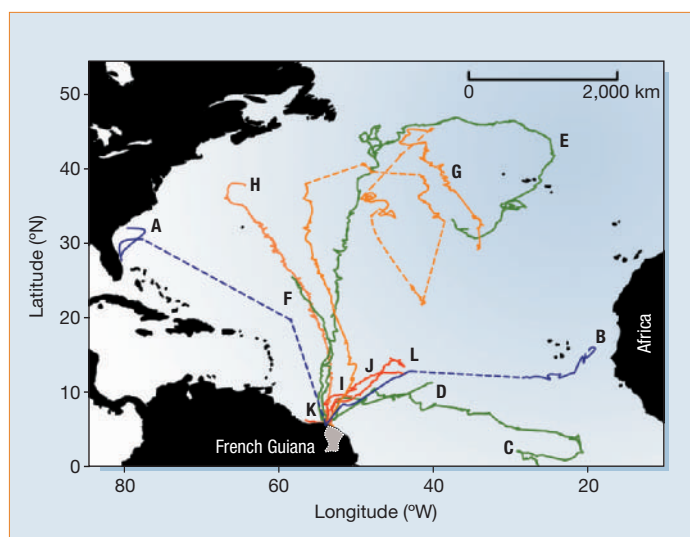


Figure 2 Reconstructed movements of 12 leatherback turtles (A–L) nesting in French Guiana and Suriname in 1999 (blue), 2000 (green), 2001 (orange) and 2002 (red). Dotted lines correspond to segments containing uncertain locations because of low transmission frequency. The leatherback turtles dispersed widely throughout the North Atlantic Ocean but mostly followed two dispersion patterns, which depended on the year: four headed north to the Gulf Stream area, where they started to move in a clockwise direction, following the general ocean circulation; six others headed east, swimming mostly against the North Equatorial Current.

2. Hays, G. C., Broderick, A. C., Godley, B. J., Luschi, P. & Nichols, W. J. *Mar. Ecol. Progr. Ser.* **262**, 305–309 (2003).
3. Morreale, S. J., Standora, E. A., Spotila, J. R. & Paladino, F. V. *Nature* **384**, 319–320 (1996).
4. Bleakney, J. S. *Can. Field Nat.* **79**, 120–128 (1965).
5. Podesta, G. P., Browder, J. A. & Hoey, J. J. *Cont. Shelf Res.* **13**, 253–277 (1993).
6. Laurs, R. M., Fielder, P. C. & Montgomery, D. R. *Deep-Sea Res.* **31**, 1085–1099 (1984).
7. Lewison, R. L., Freeman, S. A. & Crowder, L. B. *Ecol. Lett.* **7**, 221–231 (2004).
8. Fonteneau, A. *Atlas of Tropical Tuna Fisheries* (Orstom, Paris, 1997).
9. Guéguen, F. C. R. *Acad. Sci. Paris, Sciences de la Vie* **323**, 775–791 (2000).
10. Ferraroli, S., Le Maho, Y. & Georges, J.-Y. C. R. *Acad. Sci. Paris, Rapport sur la Science et la Technologie* **17**, 319–328 (2003).

Supplementary information accompanies this communication on Nature's website.

Competing financial interests: declared none.

Endangered species

Pan-Atlantic leatherback turtle movements

The overall extent of habitat use by leatherback turtles in the North Atlantic, and hence their possible interactions with longline fisheries, is unknown. Here we use long-term satellite telemetry to reveal that leatherbacks range throughout the North Atlantic, indicating that closing limited areas to longline fisheries will probably have only partial success in reducing turtle bycatch. Although turtles dive very deeply on occasion (one descended to a maximum depth of 1,230 metres, which represents the deepest dive ever recorded for a reptile), they generally restrict their diving to less than 250 metres, which increases the chance that they will encounter longline hooks.

Leatherback turtles (*Dermochelys coriacea*) are critically endangered (www.redlist.org), and their incidental capture by pelagic fisheries is a major source of mortality¹. The Atlantic is the last stronghold for leatherback turtles. Longline fishermen from the United States are mandatorily excluded from a large region of the western North Atlantic (the Northeast Distant reporting area), although other countries continue to operate there². An estimated 1.4 million hooks are deployed daily throughout the rest of the Atlantic³, an intensity of fishing that has had devastating effects on a variety of apex predators⁴. As well as trying to reduce the bycatch of leatherback turtles by using different fishing methods⁵, it is important to track their movements to identify areas that are at high risk from fisheries.

Leatherbacks routinely travel long distances and are found in the North Atlantic, far from their tropical and subtropical nesting beaches⁵. This has been shown by long-term tracking of individuals in the Pacific and Indian Oceans^{6,7} as well as by preliminary studies in the Atlantic⁸. We used satellite telemetry to record both the

horizontal and vertical (diving) movements of leatherback turtles in the North Atlantic for up to one year (for methods, see supplementary information).

We found that turtles travelled extensively throughout the Atlantic, although individuals differed in the pattern of their movements (Fig. 1a). The two turtles tracked from the Caribbean in 2002 travelled mainly eastwards: one traversed the Atlantic to within 600 km of the west African coast before returning westwards (turtle A); the other reached an area about 1,000 km from the coast of South America and remained there for several months (turtle B). Turtles leaving the Caribbean in 2003 travelled to more northerly latitudes: two travelled northwest, arriving within a few hundred kilometres of Cape Cod and Nova Scotia before turning southwards (turtles C and D, respectively); the other five travelled northeast, reaching northerly latitudes between the Azores and the United Kingdom, when some turned south (turtles E, F, G, H and I). Six of the nine tracked turtles entered the Northeast Distant area and travelled extensively inside it.

The reasons for these broad individual differences in travel routes are unknown. Ocean currents seem to play little part in driving broad-scale movements, with turtles swimming against, across and with the major current systems in the North Atlantic. Periodic residence in specific areas is probably linked to locally enhanced prey availability, as both leatherbacks and other pelagic-feeding turtle species target frontal features and mesoscale eddies^{6,9}.

Normally, more than half an individual's time was spent diving to depths below 10 m (data obtained from 3,304 individual 6-hour periods; mean percentage of time spent diving in each 6-hour period, 59.0%, s.d. = 27.6). Dives were generally within the epipelagic (near-surface) zone and over 99% of all dives were shallower than 250 m (Fig. 1b). This pattern of epipelagic diving was maintained throughout the tracking of each individual. Turtles occasionally dived very deeply and we recorded a maximum dive depth of 1,230 m. However, very deep dives were sporadic and extremely rare (of 16,767 dive profiles, 55 dives reached a maximum depth of over 400 m, 15 reached over 600 m and 6 reached over 800 m).

Our results have important implications for conservation techniques. Although closing specific areas to fishermen will help to reduce the leatherback turtle bycatch, the wide-ranging movement of individuals means that future conservation measures need to operate across the basin to the ensure survival of the species. Leatherbacks sometimes dive deeply at their nesting areas¹⁰ and deep diving should reduce their interaction with fishing gear, but our long-term telemetry

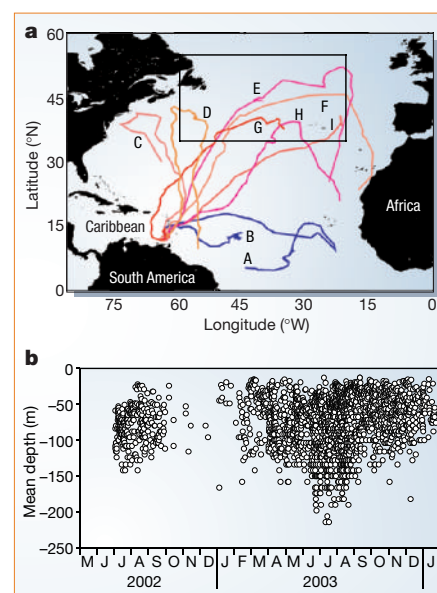


Figure 1 Movements and diving behaviour of nine leatherback turtles tracked after nesting in the Caribbean. **a**, Tracks of turtles A–I: A and B were tracked for 12 months from July 2002 until July 2003; turtles C–I were tracked for 6–8 months from May–July 2003 until January 2004. Black box outlines the Northeast Distant area, which is closed at present to longline fishing by the United States. **b**, Mean depth of dives during individual 6-hour periods, recorded from all nine turtles. For clarity, we include only those 6-hour periods during which at least 50% of the time was spent diving to over 10 metres (2,227 out of 3,304). Months indicated by initials.

results indicate that they spend most of their time diving in the epipelagic zone — which is exactly the depth-range targeted by longline fishermen¹¹.

Pan-oceanic movements and shallow diving are doubly disadvantageous, in that they both increase the interaction of leatherback turtles with longline fisheries. It is therefore crucial that new methodology and fishery management procedures be applied to reduce leatherback turtle bycatch.

Graeme C. Hays, Jonathan D. R. Houghton, Andrew E. Myers

School of Biological Sciences, Institute of Environmental Sustainability, University of Wales Swansea, Singleton Park, Swansea SA2 8PP, UK
e-mail: g.hays@swan.ac.uk

1. Spotila, J. R., Reina, R. D., Steyermark, A. C., Plotkin, P. T. & Paladino, F. V. *Nature* **405**, 529–530 (2000).
2. NOAA Northeast Distant Fishery Experiment (2001–03) <http://www.nmfs.noaa.gov/mediacenter/turtles/>
3. Lewison, R. L., Freeman, S. A. & Crowder, L. B. *Ecol. Lett.* **7**, 221–231 (2004).
4. Myers, R. A. & Worm, B. *Nature* **423**, 280–283 (2003).
5. Brongersma, J. D. *Zool. Verhandl.* **121**, 1–318 (1972).
6. Luschi, P. et al. *Proc. R. Soc. Lond. B* **270**, S129–S132 (2003).
7. Morreale, S. J., Standora, E. A., Spotila, J. R. & Paladino, F. V. *Nature* **384**, 319–320 (1996).
8. Eckert, S. A. NOAA Tech. Memo. NMFS–SEFSC **415**, 46–48 (1998).
9. Polovina, J. J. et al. *Fish. Oceanogr.* **13**, 36–51 (2004).
10. Eckert, S. A., Nellis, D. W., Eckert, K. L. & Kooyman, G. L. *Herpetologica* **42**, 381–388 (1986).
11. Yokawa, K. *Col. Vol. Sci. Pap. ICCAT* **55**, 475–479 (2003).

Supplementary information accompanies this communication on Nature's website.

Competing financial interests: declared none.

Activity-dependent homeostatic specification of transmitter expression in embryonic neurons

Laura N. Borodinsky, Cory M. Root, Julia A. Cronin, Sharon B. Sann, Xiaonan Gu & Nicholas C. Spitzer

Neurobiology Section, Division of Biological Sciences and Center for Molecular Genetics, UCSD, La Jolla, California 92093-0357, USA

Neurotransmitters are essential for interneuronal signalling, and the specification of appropriate transmitters in differentiating neurons has been related to intrinsic neuronal identity and to extrinsic signalling proteins. Here we show that altering the distinct patterns of Ca^{2+} spike activity spontaneously generated by different classes of embryonic spinal neurons *in vivo* changes the transmitter that neurons express without affecting the expression of markers of cell identity. Regulation seems to be homeostatic: suppression of activity leads to an increased number of neurons expressing excitatory transmitters and a decreased number of neurons expressing inhibitory transmitters; the reverse occurs when activity is enhanced. The imposition of specific spike frequencies *in vitro* does not affect labels of cell identity but again specifies the expression of transmitters that are inappropriate for the markers they express, during an early critical period. The results identify a new role of patterned activity in development of the central nervous system.

The determination of neuronal phenotypes is a substantial developmental challenge, given the complexity of the nervous system. The classical low-molecular-mass, peptide, gaseous and growth-factor neurotransmitters number 50 or more¹. The appearance of a particular transmitter in a given class of neurons is a crucial step in differentiation because it enables the neurons to communicate with others with which they make synaptic connections. Expression of an incorrect transmitter could isolate neurons from their normal networks. The absence of synaptic signalling could also reduce trophic support from its postsynaptic partners², leading to neuronal death.

Several classes of mechanism regulate transmitter expression. Specification by intrinsic transcription factors has been demonstrated by the ectopic expression of MNR2 or Lhx3/Is1 in the chick spinal cord, which drives inappropriate expression of the motor neuron transmitter acetylcholine in interneurons^{3,4}. In mice that are mutant for the Dbx1 transcription factor, neural progenitors from the Dbx1 domain give rise to interneurons with an inappropriate GABAergic phenotype⁵. Gain-of-function and loss-of-function experiments with homeodomain transcription factors in *Caenorhabditis elegans* and *Drosophila* lead to the misexpression of GABA and a loss of synthesis of dopamine and serotonin^{6,7}. Cytokines and neurotrophic factors can also regulate transmitter expression and can drive the expression of acetylcholine instead of noradrenaline in rat sympathetic ganglion neurons, both in culture and *in vivo*^{8–10}. Additionally, the imposition of activity can regulate the choice of neurotransmitter in cultured neurons by means of Ca^{2+} influx¹¹ and can differentially affect the regulation of transmitter expression by protein factors¹². The incidence of neurons expressing the transmitter GABA and its synthetic enzyme, glutamic acid decarboxylase (GAD), is upregulated in cultured embryonic spinal neurons by increasing the frequencies of Ca^{2+} spikes that mimic endogenous spontaneous activity^{13,14}.

Here we have examined the role of electrical activity and Ca^{2+} influx in the specification of neurotransmitter expression in the developing spinal cord of *Xenopus laevis* embryos. This is an attractive system to study because it contains only eight classes of neurons that collectively express four classical transmitters^{15,16}. Imaging neurons of four of these classes, we find they generate distinct patterns of Ca^{2+} spikes *in vivo* shortly after neural tube formation, starting before the synapse formation that enables net-

work activity¹⁷. We show that suppressing or enhancing this activity in the neural tube, either by misexpression of K^+ or Na^+ channels or with pharmacological antagonists and agonists, alters the expression of acetylcholine and glutamate without affecting markers of neuronal identity. In a homeostatic manner, spike suppression increases the incidence of excitatory transmitter expression and decreases the incidence of inhibitory transmitters. Conversely, enhancing spike production decreases the expression of excitatory transmitters and increases the expression of inhibitory transmitters. Cell cultures permit the imposition of specific spike frequencies, and we find that transmitter expression *in vitro* is dependent on Ca^{2+} spike frequency. We demonstrate transmitter release from neurons that is inappropriate for the markers they express, indicating that transmitter switches are functional. Finally we show that the effects of activity on transmitter specification are restricted to a critical period at this early stage of development.

Neuron-specific patterns of spike activity *in vivo*

To determine patterns of neuronal activity in the embryonic spinal cord, we imaged spontaneous Ca^{2+} spikes¹⁸ in dorsal sensory Rohon–Beard neurons (RB), dorsolateral interneurons (DLI), ventral motoneurons (MN) and ventral interneurons (VI) (Fig. 1a, c). We classified neurons on the dorsal surface of the neural tube as dorsomedial and dorsolateral neurons by their positions. Neurons located along the midline of the dorsal spinal cord are sensory RB neurons^{15,16}; whole-mount immunocytochemistry with antibodies against HNK-1—a membrane glycoprotein specifically expressed on RB cell bodies and processes¹⁹—confirmed this identity (Fig. 1b). We classified neurons on the ventral surface of the neural tube on the basis of the presence of coactivity. Ventral neurons that fired in concert were considered to be MN^{15,17}, whereas neurons that did not spike together with others were designated VI. The clusters of MN identified by Ca^{2+} spike coactivity (Fig. 1c and Supplementary Fig. 1) colocalized with neurons immunostained with lim-3 transcription factor in the ventral neural tube (Fig. 1d). All lim-3-immunoreactive neurons stained for choline acetyltransferase (ChAT), a generic MN marker²⁰, and vice versa (see below). Zebrafish VeLD interneurons express lim-3 and are GABA-immunoreactive^{21,22}, but we observe no GABA immunoreactivity in lim-3⁺ neurons. These

results indicate further that at the *Xenopus* developmental stages evaluated, *lim-3* is expressed only in MN²³. During a 10-h developmental period after closure of the neural tube, the incidence of spike activity increases for RB, DLI and MN but decreases for VI. The frequency patterns of spikes are different for RB (low and constant), DLI (monotonically increasing), MN (step from low to high) and VI (high throughout) (Fig. 1e, f). These observations indicated that neurons might be recognizable by signatures of activity as well as by their position and expression of molecular markers, and raised the possibility that activity drives neurotransmitter phenotype. This notion led to the hypothesis that perturbing these patterns of activity would alter the specification of neurotransmitters.

Channel overexpression alters transmitter phenotype

To examine the role of activity in neuronal differentiation, we suppressed Ca^{2+} spikes by overexpression of human inward rectifier K^+ channels (hKir2.1) and later assessed the presence of neurotransmitters immunocytochemically (Fig. 2a). Imaging embryonic neural tubes loaded with the voltage-sensitive indicator bisoxonol

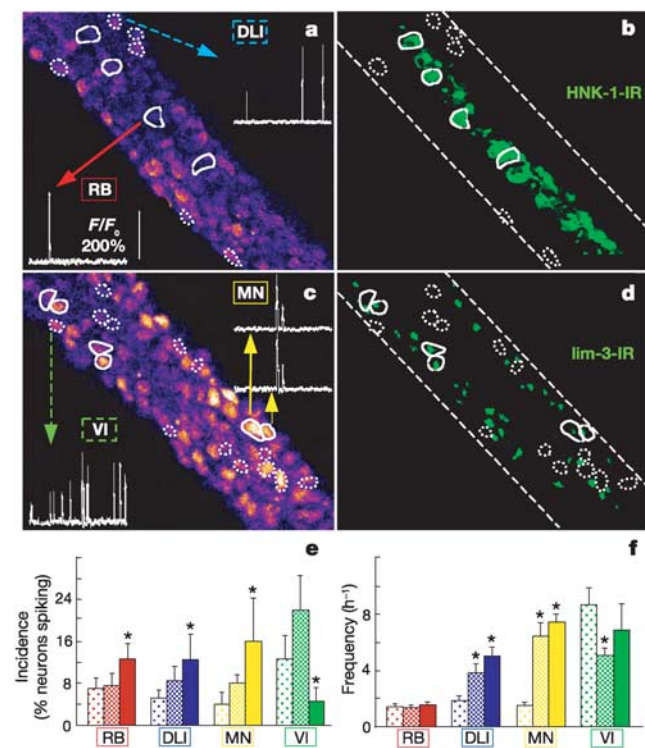


Figure 1 Ca^{2+} spike activity of four classes of neurons imaged in the embryonic spinal cord. **a**, Active Rohon–Beard neurons (RB, continuous circles) and dorsolateral interneurons (DLI, dashed circles) on the dorsal surface of a stage 23 neural tube; insets illustrate spike activity for cells indicated by arrows during a 1-h period. F/F_0 , fluorescence increase above baseline. **b**, Whole-mount immunoreactivity for HNK-1 identifies RBs on the dorsal surface of the same preparation. Dashed white lines indicate the margins of the neural tube. **c**, Coactive motor neurons (MN, continuous circles) and ventral interneurons (VI, dashed circles) on the ventral surface of a stage 24 neural tube; insets illustrate spike activity for cells indicated by arrows during a period of 1 h; scale as in **a**. **d**, Whole-mount immunoreactivity for *lim-3* identifies MNs on the ventral surface of the same preparation. Profiles of nuclei are of different sizes as the result of the through-series projection of a small stack of images and differences in nuclear orientation. **e**, Incidence of Ca^{2+} spiking (percentage active cells) for these neurons during three developmental periods. **f**, Frequency of Ca^{2+} spikes (spikes h^{-1}) for these neurons, excluding neurons that were silent during the imaging period. In **e** and **f**, $n > 10$ embryos were used for each period. Values are means $\pm$ s.e.m. and asterisks indicate significant difference from stages 20–22. Dotted columns, stages 20–22; hatched columns, stages 23–25; solid columns, stages 26–28. The neural tube is about $100 \mu\text{m}$ in diameter at these stages.

revealed that unilateral expression of hKir2.1 causes the hyperpolarization of neurons only on the ipsilateral side. Imaging with fluo-4 demonstrated ipsilateral suppression of spontaneous spikes (Fig. 2b, c). Bilateral expression hyperpolarized neurons and silenced spikes on both sides of the neural tube. Glutamate immunoreactivity (Glu-IR) and glutamate vesicular transporter immunoreactivity (VGluT-IR), and choline acetyltransferase

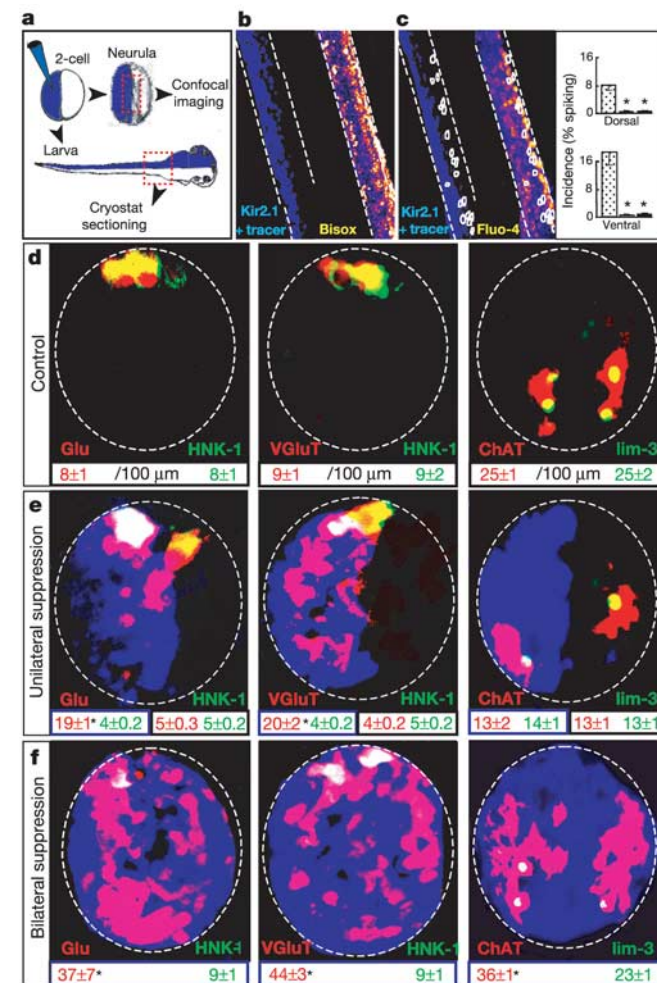


Figure 2 Suppression of spike activity *in vivo* by overexpression of inward rectifier K^+ channels increases the incidence of expression of glutamatergic and cholinergic phenotypes. **a**, Experimental design. hKir2.1 transcripts and fluorescent tracer were injected together into one or both blastomeres at the two-cell stage. Ca^{2+} imaging was performed on stage 22–26 neural-tube embryos (boxed region) and stage 40 larvae were sectioned for immunocytochemistry. **b**, Neural tube resulting from unilateral injection of transcripts plus tracer (left), loaded with bisoxonol (Bisox) to image membrane potential (right), reveals that dorsolateral neurons containing transcripts are hyperpolarized (cells are pseudocoloured blue; $n = 7$ neural tubes). White dashed lines indicate margins of the neural tube. **c**, Neural tube resulting from unilateral injection of transcripts plus tracer (left), loaded with fluo-4 acetoxymethyl ester to image spikes (middle), reveals that spikes in dorsal neurons are suppressed on the side containing transcripts (active cells are circled). The incidence of spiking is reduced in both dorsal and ventral neurons marked with tracer ($n = 15$ dorsal and ventral neural tubes). Dotted columns, control; hatched columns, Kir2.1 unilateral; solid columns, Kir2.1 bilateral. **d**, Neural-tube sections from control embryos stained for glutamate (Glu) or the vesicular glutamate transporter (VGluT) in combination with HNK-1, and choline acetyltransferase (ChAT) in combination with *lim-3*. **e**, Embryos unilaterally silenced; **f**, bilaterally silenced and stained as in **d**. White dashed ovals indicate neural-tube perimeters in this and subsequent figures. Numbers of immunoreactive neurons per $100 \mu\text{m}$ of neural tube are indicated beneath panels. For unilaterally silenced embryos these numbers are tabulated separately for each side of the neural tube. In **c–f**, values are means $\pm$ s.e.m. and asterisks indicate significantly different from control.

immunoreactivity (ChAT-IR), are normally observed only in RB and MN identified by HNK-1-IR and lim-3-IR, respectively (Fig. 2d). Strikingly, Glu-IR and VGluT-IR were present in HNK-1⁺ cells expressing hKir2.1 either unilaterally or bilaterally (Fig. 2e, f). In contrast, ChAT-IR was present in lim-3⁺ cells only after bilateral (and not unilateral) injections. Unilateral suppression might allow commissural axon projections from the contralateral side¹⁶ to prevent expansion of the cholinergic phenotype. The numbers of HNK-1⁺ and lim-3⁺ cells were not affected by either unilateral or bilateral suppression of activity, and the cellular organization of the neural tube seemed normal (Supplementary Fig. 2). These results indicate that suppression of activity might lead to the spread of glutamatergic and cholinergic phenotypes.

To test whether increases in Ca²⁺ spike frequency suppress the incidence of Glu-IR and ChAT-IR, we overexpressed voltage-gated rat brain Na⁺ channels (rNa_v2αα and rNa_v2αβ). Both the incidence of spiking neurons and spike frequency were increased as a result (Fig. 3a–c). Glu-IR and VGluT-IR were decreased after overexpression of rNa_v2ααβ both unilaterally and bilaterally, and ChAT-IR was decreased in cells after bilateral increases in spike activity (Fig. 3d, e). Unilateral expression of Na⁺ channels caused a decrease in activity in

contralateral, unlabelled VI that might result from commissural axon projections. Consistent with this level of activity there was a significant increase in ChAT-IR in lim-3⁺ cells on the contralateral side. In summary, suppression and enhancement of activity thus seem to exert opposing effects on expression of these excitatory transmitters.

Pharmacological intervention alters transmitter phenotype

To check the results of channel overexpression and to focus the manipulation of Ca²⁺ spikes on the period when this activity is manifested after closure of the neural tube, a pharmacological approach was adopted. We implanted agarose beads²⁴ immediately before the time of tube closure, loaded either with Ca²⁺ spike blockers²⁵ or with veratridine to activate Na⁺ channels and increase Ca²⁺ spike activity (Fig. 4a). The effectiveness of these agents was demonstrated by the suppression of spontaneous activity by the blockers and its enhancement by veratridine (Fig. 4b, c), with no effect of bovine serum albumin (BSA, control) when applied in the bath during imaging experiments. After implantation (Fig. 4d), Glu-IR and ChAT-IR were bilaterally increased when beads contained blockers, suppressed when beads contained veratridine (Fig. 4e, f), and unaffected when beads contained BSA. The

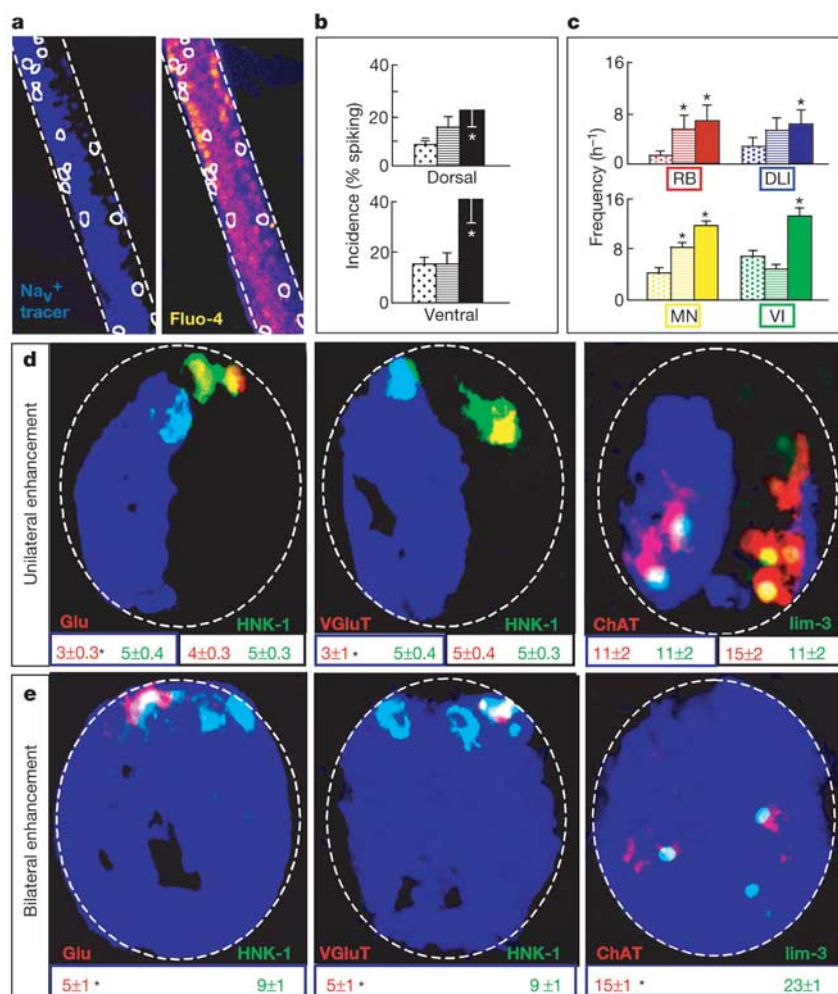


Figure 3 Enhancement of spike activity *in vivo* by overexpression of voltage-gated Na⁺ channels decreases the incidence of glutamatergic and cholinergic phenotypes. rNa_v2ααβ transcripts and fluorescent tracer were injected together into one or both blastomeres at the two-cell stage, followed by imaging and immunocytochemistry as in Fig. 2. **a**, Neural tube after unilateral injection of transcripts plus tracer (left) and fluo-4 imaging (right) shows that spike frequency is enhanced in dorsal neurons containing transcripts; spiking cells are circled. **b**, Spike incidence in dorsal and ventral neurons after unilateral or bilateral injection of transcripts and tracer. **c**, Spike frequency in RB, DLI, MN

and VI marked with tracer. In **b** and **c**, $n \geq 10$ dorsal and ventral neural tubes were analysed. Dotted columns, control; striped columns, Na_v2αα unilateral; solid columns, Na_v2αα bilateral. **d**, **e**, Embryos with unilateral and bilateral enhancement of activity, stained for glutamate or VGluT in combination with HNK-1 or for ChAT in combination with lim-3. For each condition, $n \geq 5$. For unilaterally silenced embryos these numbers are tabulated separately for each side of the neural tube. In **b–e**, values are means $\pm$ s.e.m. and asterisks indicate significantly different from control.

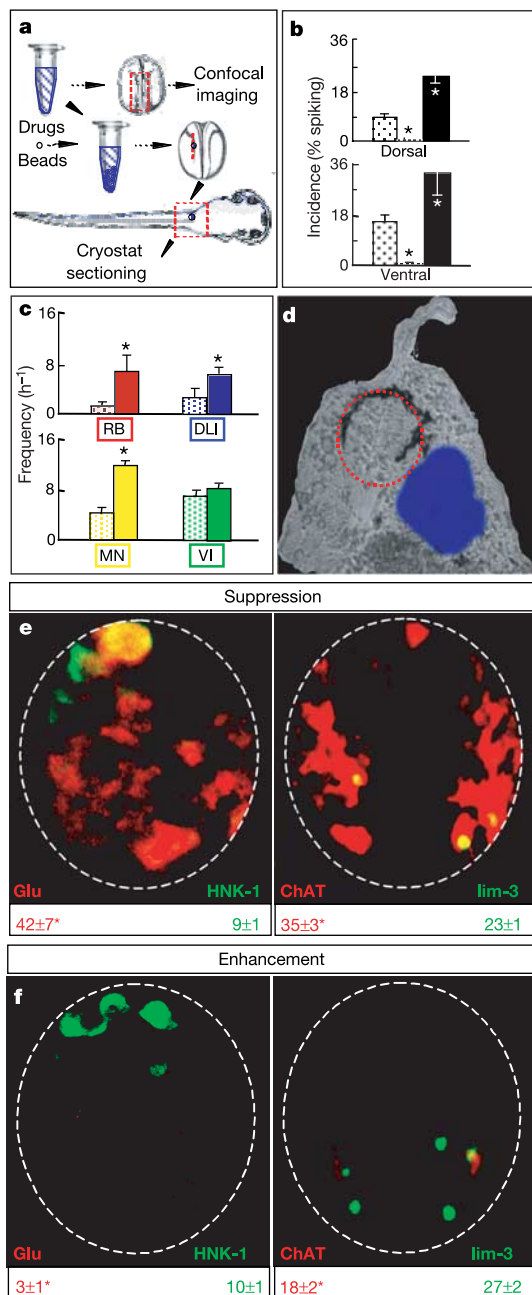


Figure 4 Pharmacological *in vivo* suppression of spikes with Ca^{2+} and Na^{+} channel blockers or enhancement of spikes with the Na^{+} channel agonist veratridine enhances or suppresses Glu-IR and ChAT-IR, respectively. **a**, Experimental design. Ca^{2+} imaging was performed on stage 22–26 embryos (boxed region) in the presence of pharmacological agents; single agarose beads loaded with these agents were implanted adjacent to the nascent neural tube at stage 17–18, and stage 40 larvae were sectioned for immunocytochemistry. **b**, Incidence of spike activity in dorsal and ventral neurons in the presence of Ca^{2+} and Na^{+} channel blockers or veratridine. Dotted columns, control; hatched columns, channel blockers; solid columns, veratridine. **c**, Spike frequency in RB, DLI, MN and VI in the presence of veratridine. Dotted columns, control; solid columns, veratridine. In **b** and **c**, $n \geq 10$ dorsal and ventral neural tubes were analysed. **d**, Bead implanted in a stage 18 embryo sectioned at stage 40. The bead is blue and the neural tube is outlined in red. **e, f**, Glu/HNK-1 staining and ChAT/lim-3 staining following implantation of beads containing blockers (e) or veratridine (f). In **e** and **f**, $n \geq 5$ embryos. Values are means $\pm$ s.e.m. and asterisks indicate significantly different from control.

difference between the effects of unilateral channel overexpression and unilateral bead implantation could be due to the diffusion of agents in the latter case. The results of these pharmacological perturbations confirm and extend the results of channel overexpression and demonstrate an inverse relationship between Ca^{2+} spike activity and the expression of excitatory transmitters.

Activity regulates transmitter phenotype homeostatically

To test the idea that transmitter specification is regulated homeostatically by Ca^{2+} spike activity, we investigated the expression of GABA and glycine. These transmitters are inhibitory for most of these embryonic *Xenopus* spinal neurons²⁶, although they are excitatory in other developing systems. Decreases in expression of Gly-IR and GABA-IR accompanied the increases in Glu-IR and ChAT-IR after bilateral suppression of activity (Fig. 5a, b). The increased incidence of excitatory transmitters seemed to lead to coexpression with inhibitory transmitters in some cases. In others, it was correlated with the disappearance of glycine and GABA, indicating the possible replacement of inhibitory by excitatory transmitters. In a reciprocal manner, the decrease in Glu-IR and ChAT-IR after bilateral enhancement of activity was accompanied by an increase in the incidence of GABA⁺ and glycine⁺ neurons relative to controls. Overexpression of rNa_v2a α β led to the appearance of RB immunoreactive for HNK-1 and GABA or glycine, in the presence or absence of Glu-IR. Similarly, we observed MN immunoreactive for lim-3 and glycine or GABA, in the presence or absence of ChAT-IR (Fig. 5c). These findings imply homeostatic regulation of transmitter expression by activity, overriding other developmental cues. Downregulation of activity seems to stimulate the production of excitatory transmitters in inhibitory neurons and to suppress the expression of inhibitory transmitters in neurons that normally express them. Upregulation of activity leads to an increased incidence of neurons expressing inhibitory transmitters and to the expression of inhibitory transmitters in neurons normally expressing excitatory transmitters.

Spike frequency regulates transmitter expression *in vitro*

To determine more precisely the effect of different Ca^{2+} spike frequencies on neurotransmitter expression, we examined neurons in low-density-dissociated cell cultures in the absence of synaptic connections. Suppression of spike activity during development *in vitro* increased the incidence of expression of Glu-IR and ChAT-IR, paralleling observations *in vivo*, without affecting the number of neurons expressing RB and MN markers of cell identity (HNK-1 and lim-3; Fig. 6a, b). We then imposed various Ca^{2+} spike frequencies on differentiating neurons and examined Glu-IR, VGLuT-IR and ChAT-IR. The expression of HNK-1 and lim-3 was constant across the range of frequencies of imposed Ca^{2+} spikes, whereas the incidence of Glu⁺/VGLuT⁺ or ChAT⁺ neurons varied inversely with frequency. Low frequencies enhanced the expression of these transmitters, and high frequencies suppressed their expression (Fig. 6c–e). The developmental progression of spike frequencies observed in the neural tube (Fig. 1f) was most effective in enhancement. These findings are consistent with results *in vivo* and provide further information about the spike activity patterns that are sufficient to regulate transmitter expression. However, transmitter expression also depends on factors related to cell identity over a wide range of the frequencies tested, because cells immunopositive for HNK-1 or lim-3 were preferentially immunoreactive for glutamate or ChAT. Activity overrides this dependence on intrinsic cell identity at the higher Ca^{2+} spike frequencies. The balance of excitatory and inhibitory neurotransmitters changed with the level of spike activity: low activity (0 mM Ca^{2+} , spikes absent) stimulated increases in excitatory and decreases in inhibitory transmitters. In contrast, high activity (2 mM Ca^{2+} , spikes present at 10 h⁻¹)²⁵ stimulated increases in inhibitory and decreases in excitatory transmitters (Fig. 6f, g). These results are consistent with the

increased incidence of expression of GABA and xGAD 67 with increasing spike frequency¹⁴ and indicate that activity-dependent regulation of transmitter expression might be cell autonomous.

Demonstration of transmitter release

Given the changes in incidence of Glu-IR and ChAT-IR when

neurons are subjected to different patterns of spike activity, we then tested whether neurons are able to release these transmitters. When acetylcholine-receptor-expressing myoballs were brought into contact with neuronal growth cones, we recorded spontaneous synaptic currents (SSCs) that were blocked by curare but not by 6-cyano-7-nitroquinoxaline-2,3-dione (CNQX). We used blastomere

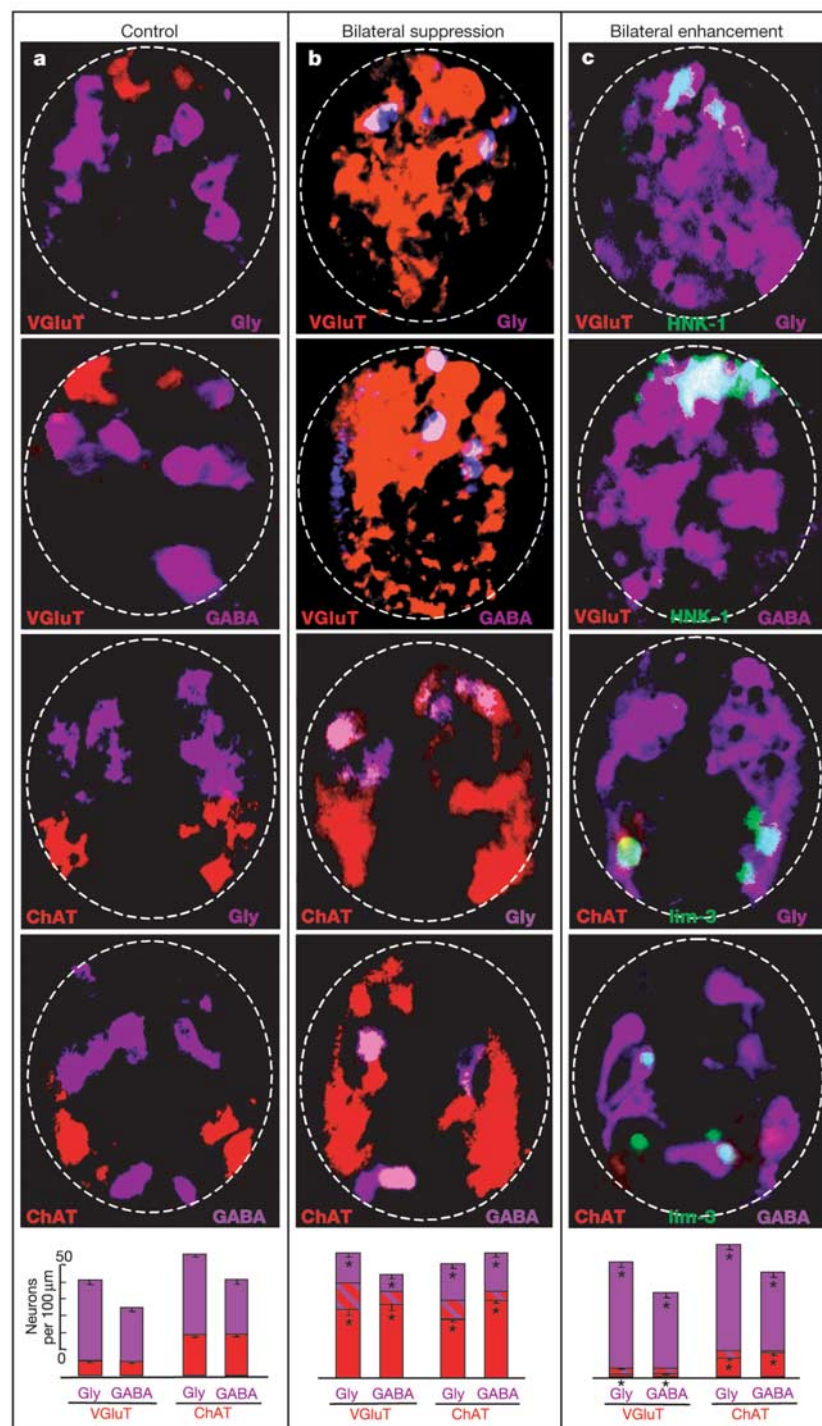


Figure 5 Suppression or enhancement of spike activity *in vivo* causes homeostatic superposition or replacement of one transmitter with another. **a**, Controls doubly stained for VGLuT or ChAT (red, excitatory) plus glycine or GABA (purple, inhibitory) illustrate the normal distribution of immunoreactivity. **b**, Embryos in which spike activity was bilaterally suppressed by expression of hKir2.1, stained as in **a**. Pink indicates coexpression of excitatory and inhibitory transmitters. **c**, Embryos in which spike activity was bilaterally enhanced by expression of rNa_v2αβ, stained for glutamate or ChAT and the respective

marker of cell identity (HNK-1 or lim-3, green), plus glycine or GABA. Light blue denotes coexpression of marker and inhibitory transmitter; white shows coexpression of marker and both excitatory and inhibitory transmitters. Bottom, number of neurons immunoreactive for different transmitters per 100 μm of neural tube, means ± s.e.m.; $n \geq 5$ embryos for each condition. In **b** and **c**, hatched columns indicate number of neurons coexpressing both neurotransmitters. Asterisks indicate significantly different from control.

injection of rGluR2 messenger RNA to generate myoballs expressing glutamate receptors; when these myoballs were brought in contact with growth cones, we recorded SSCs that were blocked by CNQX but not by curare. The absence of cholinergic SSCs in this condition might have been due to downregulation of AChR by overexpression of GluR2. The percentages of neurons generating glutamatergic and cholinergic SSCs when cultured in medium containing 0 or 2 mM Ca^{2+} (Fig. 7; $n \geq 10$) were not distinguishable from the incidence of Glu-IR and ChAT-IR (Fig. 6a, b), indicating that these altered phenotypes are functionally expressed. A parallel analysis of evoked synaptic currents elicited by the electrical stimulation of neurons yielded similar percentages (data not shown). The results can account for the high incidence of neuromuscular junctions in nerve-muscle cocultures²⁷.

Critical periods for regulation by Ca^{2+} spikes

Regulatory roles of activity during neuronal development are often confined to critical periods^{28,29}. To determine whether this is true of transmitter expression, we suppressed or imposed spikes for various times *in vitro* to identify the minimal stimulation required to achieve changes in incidence of transmitter expression. Five-hour periods of perturbation beginning near the time at which Ca^{2+} spikes are first expressed are sufficient to increase or decrease the incidence of Glu-IR or ChAT-IR to the same extent as that attained

after 12 h. These seem to be critical periods, because the reversal of stimulation or deprivation after 5 h is ineffective in reversing the changes during the periods tested (Fig. 8).

Discussion

Our results indicate that specific patterns of Ca^{2+} spike activity are necessary for normal expression of neurotransmitters in neurons in the embryonic spinal cord, because disrupting these patterns alters transmitter expression. Neuronal death or neurogenesis seems unlikely to contribute to these results because the numbers of RB and MN identified by HNK-1 and lim-3 do not change *in vivo* or *in vitro*. Ca^{2+} spike activity might also be sufficient to drive the expression of particular transmitters, because patterns of stimulation that more closely parallel the *in vivo* activity of RB or MN generate an increasing proportion of neurons expressing glutamate or ChAT *in vitro*. However, a large percentage of these neurons are immunoreactive for both glutamate and ChAT, and the mechanism by which neurons are led to express a single classical transmitter during normal development *in vivo* remains to be determined. Candidates include regulation by transcription factors, more subtle distinctions among patterns of spontaneous activity, the presence of synaptic activity, and signal transduction by means of protein factors. Natural changes in spike activity patterns could have a role in the restriction of the extensive early expression of GAD and

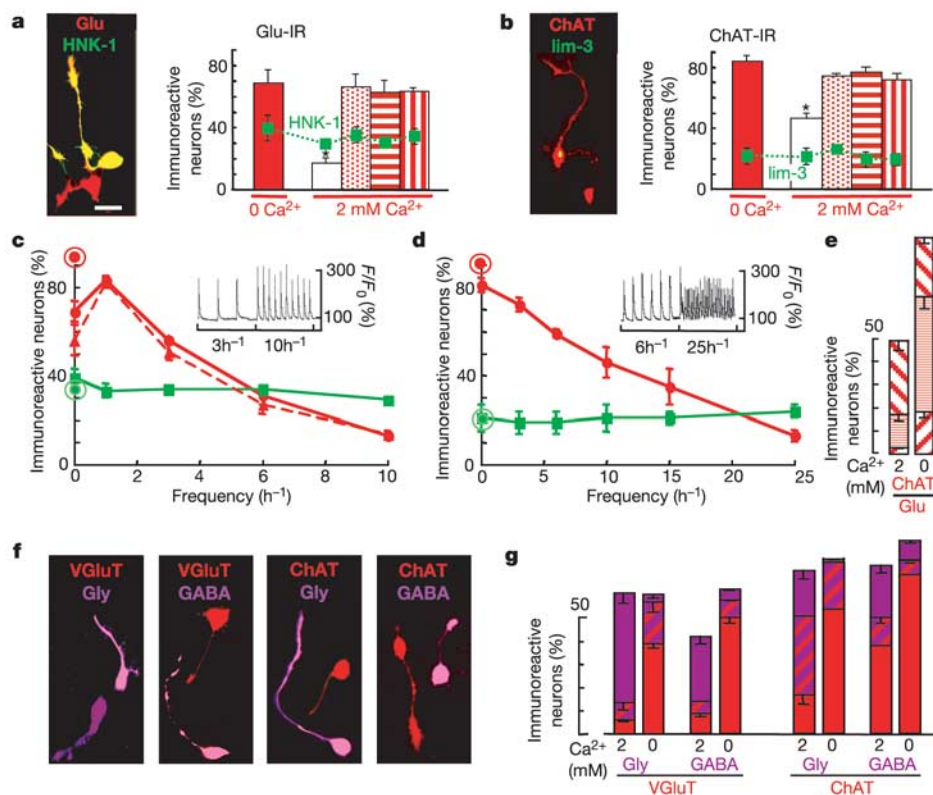


Figure 6 Regulation of spike frequency *in vitro* drives novel expression of neurotransmitters. **a**, Examples of Glu⁺/HNK-1⁺ and Glu⁺/HNK-1⁻ phenotypes, and HNK-1-IR (green) and Glu-IR in 2 mM Ca^{2+} (white) or after suppression of spike activity with 0 mM Ca^{2+} (solid tint) or 2 mM Ca^{2+} culture medium with bis-(*o*-aminophenoxy)ethane-*N,N,N',N'*-tetraacetic acid acetoxymethyl ester (BAPTA-AM, dotted tint), Ca^{2+} channel blockers (horizontal hatching) or hKir2.1 expression (vertical hatching). Scale bar, 25 μm . **b**, Examples of ChAT⁺/lim-3⁺ and ChAT⁺/lim-3⁻ phenotypes, and lim-3-IR (green) and ChAT-IR in 2 mM Ca^{2+} (white) or after suppression of spike activity as in **a**. Asterisks indicate significantly different from 0 mM Ca^{2+} . **c**, Staining for Glu (circles), VGLUT (triangles) and HNK-1 (squares) as a function of frequency of imposed Ca^{2+} spikes. HNK-1⁺ neurons are Glu⁺/VGLUT⁺ at frequencies of 6 h⁻¹ or lower. Bullseyes indicate the effect of stimulation with the RB pattern of spikes (Fig. 1f). Inset, examples of simulated spikes at 3 and 10 h⁻¹.

d, Staining of ChAT (circles) and lim-3 (squares) as a function of frequency of imposed Ca^{2+} spikes. lim-3⁺ neurons are ChAT⁺ at frequencies of 15 h⁻¹ or lower. Bullseyes indicate the effect of stimulation with the MN pattern of spikes (Fig. 1f). Inset, simulated spikes at 6 and 25 h⁻¹. **e**, Proportions of Glu⁺/ChAT⁺ doubly stained neurons when cultured in 2 mM Ca^{2+} or 0 mM Ca^{2+} medium (indicated by 2 or 0, respectively). The percentage of neurons doubly labelled for glutamate and ChAT is consistent with predictions from **a-d**. **f**, Neurons doubly labelled for excitatory (red) and inhibitory (purple) transmitters; pink indicates coexpression. **g**, Incidence of neurons expressing excitatory and inhibitory transmitters when grown in 2 mM Ca^{2+} or 0 mM Ca^{2+} culture medium (indicated by 2 or 0, respectively). For **a-e** and **g**, values are means $\pm$ s.e.m. for at least five cultures containing $n \geq 40$ neurons.

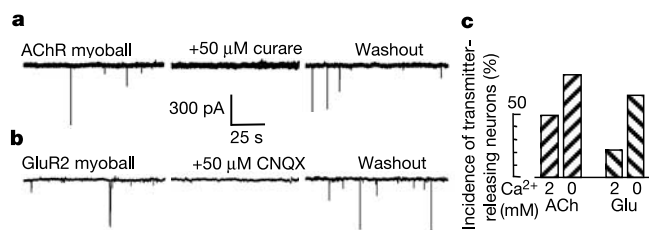


Figure 7 Neurotransmitters expressed after alterations in Ca^{2+} spike activity are functionally released. **a**, Whole-cell recording from a myoball expressing native AChR manipulated in front of the growth cone of a neuron grown in 2 mM Ca^{2+} culture medium, illustrating control SSCs in the absence and presence of curare. **b**, Similar recording from a myoball expressing rGluR2 in front of another growth cone of a neuron grown in 0 mM Ca^{2+} medium, in the absence and presence of CNQX. **c**, Percentage of neurons generating cholinergic (ACh) or glutamatergic (Glu) SSCs when cultured in 2 mM Ca^{2+} or 0 mM Ca^{2+} medium (indicated by 2 or 0, respectively) and tested with control myoballs or myoballs expressing GluR2; all recordings were made in the presence of 2 mM Ca^{2+} ($n > 9$ for each condition).

GABA³⁰. The coexpression of several transmitters in single embryonic neurons after particular patterns of activity might provide the basis for transmitter coexpression in mature neurons³¹.

Shifts in transmitter expression in response to decreases or increases in spike activity seem to be directed towards the homeostasis of network excitability in the nervous system³². This process is conceptually similar to the homeostatic resetting of neuronal excitability and synaptic strengths that provides an important balance to hebbian plasticity^{33,34} and to the homeostatic plasticity of excitatory networks that renders spontaneous output resistant to disruption of connectivity^{35,36}. It was thus surprising to observe this process in neurons in dissociated cell culture in the absence of synaptic connections, in turn implying that the feedback loop might be intracellular. This regulatory programme indicates that abnormal electrical activity could regulate neurotransmitter expression. Although high frequencies of Ca^{2+} spikes generated in paediatric epilepsy can be detrimental to brain development, seizure activity seems to regulate the expression of neurotransmitter homeostatically to suppress this activity. Mossy fibres normally generate excitatory glutamatergic postsynaptic potentials on hippocampal CA3 pyramidal cells, but generate inhibitory GABAergic postsynaptic potentials after kindled seizures³⁷.

The effects of activity on transmitter specification probably operate through dynamic interplay with the molecular context provided by transcription factors, but the rules that govern this interaction are still unclear. For example, are the peptide, gaseous and growth-factor transmitters specified in the same way as the classical transmitters, or are these other categories of transmitters regulated differently? What is the range of transmitters that can be expressed in an activity-dependent manner? Can transmitters that are not normally expressed be synthesized in spinal neurons in response to particular patterns of activity, or are switches restricted to the population of classical transmitters normally present in the network? Additionally, it will be of interest to ascertain whether the time frames of action of transcription factors and activity are coextensive or the effects of one ultimately outweigh the other.

We propose a model in which the hierarchical expression of transcription factors defines fields of cells^{38–40} that express constellations of ion channels. These channels produce patterns of activity that are modulated by protein factors. This modulated activity further engages transcription factors that stipulate transmitter expression. The transmitter that is specified then depends on the transcription factors expressed by the postmitotic neuron, the appropriate type of activity, interaction with signalling proteins, and further transcriptional regulation. This scheme enables the integration of genetic coding with signals that stimulate spike activity or the secretion of factors. A consequence of this proposal is that knockouts of neuronal

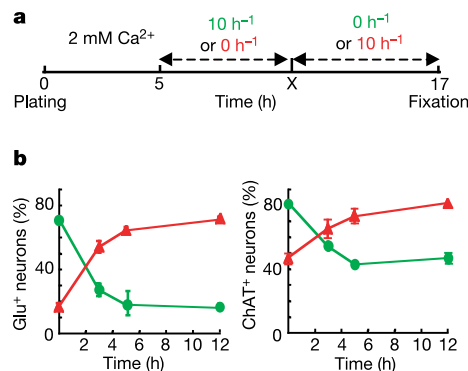


Figure 8 Critical periods for Ca^{2+} spike-dependent regulation of transmitter expression *in vitro*. **a**, Experimental design. To test the existence of critical periods, growth in 2 mM Ca^{2+} for 5 h after plating was followed by exchange of the culture medium for one containing 0 mM Ca^{2+} (deprivation) or by the imposition of spikes at 10 h^{-1} (stimulation) for various periods. Treatment was reversed after X hours, and cultures were fixed and stained for glutamate or ChAT at 17 h in culture. **b**, Glutamate staining as a function of the period of initial stimulation (circles, spike frequency 10 h^{-1}) or deprivation (triangles, spike frequency 0 h^{-1}), and ChAT staining as a function of the period of initial stimulation or deprivation. In both parts of **b**, $n \geq 5$ cultures were used containing $n \geq 40$ neurons; values are means $\pm$ s.e.m.

class-specific transcription factors might not lead to transmitter switches unless they are involved in programming electrical activity⁴¹. Moreover, the robustness of transmitter phenotype in cells dissociated and grown in culture⁴² implies the preservation of patterns of activity as observed in this study. Consistent with this model is the observation that misexpression of a lim-class homeobox gene in epidermal cells leads to ectopic expression of a putative voltage-dependent Na^+ channel gene normally expressed in motor neurons⁴³, and neurotrophins modulate Ca^{2+} and K^+ channels⁴⁴. The signal transduction cascades by which these low frequencies of spikes exert these effects are likely to engage RNA synthesis^{14,39} by means of activity-dependent transcription factors^{45,46}.

Our results lead to several predictions. On the one hand, activity-dependent changes in transmitter expression can be expected to affect axon guidance. Growth cones of developing *Xenopus* neurons release transmitter spontaneously⁴⁷ and turn in response to acetylcholine, glutamate and GABA^{48,49}. In agreement with this view is the finding that the depletion of transmitters *in vivo* alters axon outgrowth⁵⁰. On the other hand, neurons rerouted by novel neurotransmitter expression can be expected to make synaptic connections with novel postsynaptic partners. Embryonic *Xenopus* neurons express multiple classes of transmitter receptors²⁶, providing key components necessary for the formation of functional connections. Alternatively, the expression of postsynaptic receptors might be regulated homeostatically in parallel with the regulation of transmitter expression. However, mismatches between presynaptically released transmitter and postsynaptically expressed receptors seem likely to arise. It will be of interest to determine whether the resulting neuronal networks promote behavioural homeostasis. □

Methods

Imaging

Neural tubes dissected from three embryonic epochs¹⁸ and dissociated cell cultures prepared from neural-plate-stage embryos²⁹ were loaded with 5 μM fluo-4 acetoxymethyl ester or 1 μM bisoxonol, and images were acquired at 0.2 Hz for 1-h periods with a BioRad MRC1024 laser confocal system. Spikes were stimulated at different frequencies *in vitro* by culturing neurons in 250 μl Ca^{2+} -free saline medium by using a volume reducer and continuously superfusing them with this medium at 2.5 ml min^{-1} . The composition of saline was automatically switched for 15–20 s by computer-controlled solenoid valves (General Valve Corp.) to a solution containing 100 mM KCl and 2 mM Ca^{2+} .

Molecular biology and pharmacology

hKir2.1, rNav2a α 3 and rGluR2 were gifts from E. Marban, W. Catterall and S. Heinemann. The genes were subcloned into a Bluescript vector and complementary DNA was

transcribed with the mMessage mMachine (Ambion). Capped RNA (5–10 nl of a 0.01–0.1 mg ml⁻¹ RNA solution in 10% MMR, 6% Ficoll) was co-injected with a Cascade blue or Rhodamine red 30-kDa dextran (30 mg ml⁻¹) into one or both blastomeres at the two-cell stage with the use of a picospritzer (Picospritzer III; Parker Instrumentation). Control injections consisted of fluorescent dextran alone. Agarose beads (80 µm; BioRad) were loaded for 1 h with a solution containing 200 nM calicudine (Calbiochem), 10 µM GVIA ω-conotoxin, 10 µM flunarizine and 10 µg/ml tetrodotoxin, or 1 mM veratridine, or 1% BSA (all from Sigma) before implantation. The effect of most of these agents on spike activity was tested at one-tenth of these concentrations; veratridine was tested at one-thousandth.

Immunocytochemistry

Embryos were fixed in 4% paraformaldehyde, 0.1% glutaraldehyde, phosphate-buffered saline (pH 7.4) for 2 h at 4 °C, soaked in 30% sucrose for 2.5 h, and embedded in OCT compound (Tissue-Tek, Fisher Scientific). Frozen sections 10 µm in thickness were made over a 400-µm region of the neural tube starting about 400 µm posterior to the back of the eyes for control and channel-misexpression embryos. In bead-implanted embryos the bead was inserted between the neural tube and myotomes about 400 µm behind the eye primordium and sectioned from 100 µm anterior to 100 µm posterior to the bead. Cultures were fixed for 5–10 min with the same fixative. Slides and cultures were incubated in a blocking solution of 1% goat serum, fish gelatin or BSA for 0.5 h at 20 °C, followed by incubation overnight at 4 °C with primary antibodies, and incubation for 2 h with fluorescently tagged secondary antibodies at 20 °C. Immunoreactivity was examined on a Zeiss Photoscope with a 40X oil-immersion objective (or 20X and 40X water-immersion objectives for cultures) using a Xenon arc lamp attenuated by neutral density filters, and the appropriate excitation and emission filters for Cascade Blue, Alexa 488 and Alexa 594 fluorophores. Images were acquired and analysed with Metamorph software (Universal Imaging Corp.). Each phenotype was scored in 20–30 consecutive sections or in cultures from at least five embryos. Immunoreactivities are presented in saturated colour to clarify the distinction between positive and negative cells and to normalize image intensities. Differences from control were considered significant at $P < 0.05$ (Student's *t*-test). Antibodies were from Affinity Bioreagents, Calbiochem, Chemicon and Sigma.

Electrophysiology

Cultured neurons and myocytes (myoballs) were observed with an upright compound microscope and a 40X water-immersion objective that allowed the positioning of electrodes with phase-contrast optics. Whole-cell recording¹³ was used to record SSCs. A Dell Dimension 4100 computer with Axon Instruments (PCLamp 8.1) software and data interfaces was used for the acquisition and analysis of currents. Electrodes were pulled from borosilicate capillaries and had resistances of 3–5 MΩ; they were filled with 100 mM KCl, 10 mM EGTA, 10 mM HEPES, pH 7.4. A perfusion system allowing rapid exchange of the bathing medium was used to achieve solution changes (2 ml min⁻¹). The receptor dependence of currents was determined by adding various drugs.

Received 6 November 2003; accepted 29 March 2004; doi:10.1038/nature02518.

- Snyder, S. H. & Ferris, C. D. Novel neurotransmitters and their neuropsychiatric relevance. *Am. J. Psychiatry* **157**, 1738–1751 (2000).
- Zafra, F., Hengeler, B., Leibrock, J., Thoenen, H. & Lindholm, D. Activity dependent regulation of BDNF and NGF mRNAs in the rat hippocampus is mediated by non-NMDA glutamate receptors. *EMBO J.* **9**, 3545–3550 (1990).
- Tanabe, Y., William, C. & Jessell, T. M. Specification of motor neuron identity by the MNR2 homeodomain protein. *Cell* **95**, 67–80 (1998).
- Thaler, J. P., Lee, S. K., Jurata, L. W., Gill, G. N. & Pfaff, S. L. LIM factor Lhx3 contributes to the specification of motor neuron and interneuron identity through cell-type-specific protein-protein interactions. *Cell* **110**, 237–249 (2002).
- Pierani, A. *et al.* Control of interneuron fate in the developing spinal cord by the progenitor homeodomain protein Dbx1. *Neuron* **29**, 367–384 (2001).
- Jin, Y., Hoskins, R. & Horvitz, H. R. Control of type-D GABAergic neuron differentiation by *C. elegans* UNC-30 homeodomain protein. *Nature* **372**, 780–783 (1994).
- Thor, S. & Thomas, J. B. The *Drosophila* islet gene governs axon pathfinding and neurotransmitter identity. *Neuron* **18**, 397–409 (1997).
- Furshpan, E. J., MacLeish, P. R., O'Lague, P. H. & Potter, D. D. Chemical transmission between rat sympathetic neurons and cardiac myocytes developing in microcultures: evidence for cholinergic, adrenergic, and dual-function neurons. *Proc. Natl Acad. Sci. USA* **73**, 4225–4229 (1976).
- Landis, S. C. & Keefe, D. Evidence for neurotransmitter plasticity *in vivo*: developmental changes in properties of cholinergic sympathetic neurons. *Dev. Biol.* **98**, 349–372 (1983).
- Nawa, H. & Patterson, P. H. Separation and partial characterization of neuropeptide-inducing factors in heart cell conditioned medium. *Neuron* **4**, 269–277 (1990).
- Walickie, P. A. & Patterson, P. H. On the role of Ca²⁺ in the transmitter choice made by cultured sympathetic neurons. *J. Neurosci.* **1**, 343–350 (1981).
- Rao, M. S., Tyrrell, S., Landis, S. C. & Patterson, P. H. Effects of ciliary neurotrophic factor (CNTF) and depolarization on neuropeptide expression in cultured sympathetic neurons. *Dev. Biol.* **150**, 281–293 (1992).
- Gu, X. & Spitzer, N. C. Distinct aspects of neuronal differentiation encoded by frequency of spontaneous Ca²⁺ transients. *Nature* **375**, 784–787 (1995).
- Watt, S. D., Gu, X., Smith, R. D. & Spitzer, N. C. Specific frequencies of spontaneous Ca²⁺ transients upregulate GAD 67 transcripts in embryonic spinal neurons. *Mol. Cell. Neurosci.* **16**, 376–387 (2000).
- Hartenstein, V. Early pattern of neuronal differentiation in the *Xenopus* embryonic brainstem and spinal cord. *J. Comp. Neurol.* **328**, 213–231 (1993).
- Li, W. C. *et al.* Defining classes of spinal interneuron and their axonal projections in hatchling *Xenopus laevis* tadpoles. *J. Comp. Neurol.* **441**, 248–265 (2001).
- Hanson, M. G. & Landmesser, L. T. Characterization of the circuits that generate spontaneous episodes of activity in the early embryonic mouse spinal cord. *J. Neurosci.* **23**, 587–600 (2003).

- Gu, X., Olson, E. C. & Spitzer, N. C. Spontaneous neuronal calcium spikes and waves during early differentiation. *J. Neurosci.* **14**, 6325–6335 (1994).
- Somasekhar, T. & Nordlander, R. H. Selective early innervation of a subset of epidermal cells in *Xenopus* may be mediated by chondroitin sulfate proteoglycans. *Dev. Brain Res.* **99**, 208–215 (1997).
- Arber, S. *et al.* Requirement for the homeobox gene Hb9 in the consolidation of motor neuron identity. *Neuron* **23**, 659–674 (1999).
- Bernhardt, R. R., Patel, C. K., Wilson, S. W. & Kuwada, J. Y. Axonal trajectories and distribution of GABAergic spinal neurons in wildtype and mutant zebrafish lacking floor plate cells. *J. Comp. Neurol.* **326**, 263–272 (1992).
- Appel, B. *et al.* Motoneuron fate specification revealed by patterned LIM homeobox gene expression in embryonic zebrafish. *Development* **121**, 4117–4125 (1995).
- Taira, M., Hayes, W. P., Otani, H. & Dawid, I. B. Expression of LIM class homeobox gene Xlim-3 in *Xenopus* development is limited to neural and neuroendocrine tissues. *Dev. Biol.* **159**, 245–256 (1993).
- Schneider, C., Wicht, H., Enderich, J., Wegner, M. & Rohrer, H. Bone morphogenetic proteins are required *in vivo* for the generation of sympathetic neurons. *Neuron* **24**, 861–867 (1999).
- Gorbanova, Y. V. & Spitzer, N. C. Dynamic interactions of cyclic AMP transients and spontaneous Ca²⁺ spikes. *Nature* **418**, 93–96 (2002).
- Bixby, J. L. & Spitzer, N. C. The appearance and development of neurotransmitter sensitivity in *Xenopus* embryonic spinal neurons *in vitro*. *J. Physiol. (Lond.)* **353**, 143–155 (1984).
- Henderson, L. P., Smith, M. A. & Spitzer, N. C. The absence of calcium blocks impulse-evoked release of acetylcholine but not de novo formation of functional neuromuscular synaptic contacts in culture. *J. Neurosci.* **4**, 3140–3150 (1984).
- Hubel, D. H. & Wiesel, T. N. The period of susceptibility to the physiological effects of unilateral eye closure in kittens. *J. Physiol. (Lond.)* **206**, 419–436 (1970).
- Ribera, A. B. & Spitzer, N. C. A critical period of transcription required for differentiation of the action potential of spinal neurons. *Neuron* **2**, 1055–1062 (1989).
- Somogyi, R., Wen, X., Ma, W. & Barker, J. L. Developmental kinetics of GAD family mRNAs parallel neurogenesis in the rat spinal cord. *J. Neurosci.* **15**, 2575–2591 (1995).
- Jonas, P., Bischofberger, J. & Sandkuhler, J. Corelease of two fast neurotransmitters at a central synapse. *Science* **281**, 419–424 (1998).
- Berridge, M. J., Bootman, M. D. & Roderick, H. L. Calcium signalling: dynamics, homeostasis and remodelling. *Nature Rev. Mol. Cell Biol.* **4**, 517–529 (2003).
- Turrigiano, G. G. & Nelson, S. B. Hebb and homeostasis in neuronal plasticity. *Curr. Opin. Neurobiol.* **10**, 358–364 (2000).
- Marek, K. W. *et al.* A genetic analysis of synaptic development: pre- and postsynaptic dCBP control transmitter release at the *Drosophila* NMJ. *Neuron* **25**, 537–547 (2000).
- Chub, N. & O'Donovan, M. J. Blockade and recovery of spontaneous rhythmic activity after application of neurotransmitter antagonists to spinal networks of the chick embryo. *J. Neurosci.* **18**, 294–306 (1998).
- Milner, L. D. & Landmesser, L. T. Cholinergic and GABAergic inputs drive patterned spontaneous motoneuron activity before target contact. *J. Neurosci.* **19**, 3007–3022 (1999).
- Gutierrez, R. *et al.* Plasticity of the GABAergic phenotype of the 'glutamatergic' granule cells of the rat dentate gyrus. *J. Neurosci.* **23**, 5594–5598 (2003).
- Briscoe, J., Pierani, A., Jessell, T. M. & Ericson, J. A homeodomain protein code specifies progenitor cell identity and neuronal fate in the ventral neural tube. *Cell* **101**, 435–445 (2000).
- Brosenitsch, T. A. & Katz, D. M. Expression of Phox2 transcription factors and induction of the dopaminergic phenotype in primary sensory neurons. *Mol. Cell. Neurosci.* **20**, 447–457 (2002).
- Groves, A. K. *et al.* Differential regulation of transcription factor gene expression and phenotypic markers in developing sympathetic neurons. *Development* **121**, 887–901 (1995).
- Thaler, J. *et al.* Active suppression of interneuron programs within developing motor neurons revealed by analysis of homeodomain factor HB9. *Neuron* **23**, 675–687 (1999).
- Wichterle, H., Lieberam, L., Porter, J. A. & Jessell, T. M. Directed differentiation of embryonic stem cells into motor neurons. *Cell* **110**, 385–397 (2002).
- Okada, T., Katsuyama, Y., Ono, F. & Okamura, Y. The development of three identified motor neurons in the larva of an ascidian, *Halocynthia roretzi*. *Dev. Biol.* **244**, 278–292 (2002).
- DeFazio, R. A., Pong, K., Knusel, B. & Walsh, J. P. Neurotrophin-4/5 promotes dendritic outgrowth and calcium currents in cultured mesencephalic dopamine neurons. *Neuroscience* **99**, 297–304 (2000).
- Li, W., Llopis, J., Whitney, M., Zlokarnik, G. & Tsien, R. Y. Cell-permeant caged InsP3 ester shows that Ca²⁺ spike frequency can optimize gene expression. *Nature* **392**, 936–941 (1998).
- West, A. E., Griffith, E. C. & Greenberg, M. E. Regulation of transcription factors by neuronal activity. *Nature Rev. Neurosci.* **3**, 921–931 (2002).
- Young, S. H. & Poo, M. M. Spontaneous release of transmitter from growth cones of embryonic neurones. *Nature* **305**, 634–637 (1983).
- Zheng, J. Q., Felder, M., Connor, J. A. & Poo, M. M. Turning of nerve growth cones induced by neurotransmitters. *Nature* **368**, 140–144 (1994).
- Xiang, Y. *et al.* Nerve growth cone guidance mediated by G protein-coupled receptors. *Nature Neurosci.* **5**, 843–848 (2002).
- Diefenbach, T. J., Sloley, B. D. & Goldberg, J. I. Neurite branch development of an identified serotonergic neuron from embryonic *Helisoma*: evidence for autoregulation by serotonin. *Dev. Biol.* **167**, 282–293 (1995).

Supplementary Information accompanies the paper on www.nature.com/nature.

Acknowledgements We thank D. Berg, D. Feldman, M. Feller and A. Ghosh for comments on the manuscript, T. Jessell for discussions, and I-T. Hsieh for technical support. This work was supported by a grant to N.C.S. from the National Institutes of Health. S.B.S. was supported by the NSF and Merck.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to L.N.B. (lborodin@biomail.ucsd.edu).

The structure of the high-energy spin excitations in a high-transition-temperature superconductor

S. M. Hayden¹, H. A. Mook², Pengcheng Dai^{2,3}, T. G. Perring⁴ & F. Doğan⁵

¹H. H. Wills Physics Laboratory, University of Bristol, Bristol BS8 1TL, UK

²Condensed Matter Sciences Division, Oak Ridge National Laboratory, Oak Ridge, Tennessee 37831-6393, USA

³Department of Physics and Astronomy, The University of Tennessee, Knoxville, Tennessee 37996-1200, USA

⁴ISIS Facility, Rutherford Appleton Laboratory, Chilton, Oxon OX11 0QX, UK

⁵Department of Ceramic Engineering, University of Missouri-Rolla, Rolla, Missouri 65409-0330, USA

In conventional superconductors, lattice vibrations (phonons) mediate the attraction between electrons that is responsible for superconductivity¹. The high transition temperatures (high- T_c) of the copper oxide superconductors has led to collective spin excitations being proposed as the mediating excitations in these materials². The mediating excitations must be strongly coupled to the conduction electrons, have energy greater than the pairing energy, and be present at T_c . The most obvious feature in the magnetic excitations of high- T_c superconductors such as $\text{YBa}_2\text{Cu}_3\text{O}_{6+x}$ is the so-called ‘resonance’^{3–6}. Although the resonance may be strongly coupled to the superconductivity^{3–8}, it is unlikely to be the main cause, because it has not been found in the $\text{La}_{2-x}(\text{Ba,Sr})_x\text{CuO}_4$ family and is not universally present in $\text{Bi}_2\text{Sr}_2\text{CaCu}_2\text{O}_{8+\delta}$ (ref. 9). Here we use inelastic neutron scattering to characterize possible mediating excitations at higher energies in $\text{YBa}_2\text{Cu}_3\text{O}_{6.6}$. We observe a square-shaped continuum of excitations peaked at incommensurate positions. These excitations have energies greater than the superconducting pairing energy, are present at T_c , and have spectral weight far exceeding that of the ‘resonance’. The discovery of similar excitations in $\text{La}_{2-x}\text{Ba}_x\text{CuO}_4$ (ref. 10) suggests that they are a general property of the copper oxides, and a candidate for mediating the electron pairing.

In a conventional superconductor, the modification of the phonons¹¹ on entering the superconducting state indicates the presence of strong electron–phonon coupling. The dramatic changes seen in the magnetic excitation spectrum^{3–6,12–15} as copper oxide superconductors enter the superconducting state points to a strong coupling of the magnetic excitations and superconductivity. The magnetism in superconducting copper oxides arises from the incomplete outer shells of the copper ions and fluctuations in these unpaired spins give rise to collective magnetic excitations, which are coherent over many lattice sites. We use neutron scattering to probe the wavevector ($\mathbf{Q}$) and energy (E) dependence of the spin fluctuations in $\text{YBa}_2\text{Cu}_3\text{O}_{6.6}$. To label the two-dimensional reciprocal space which corresponds to the CuO_2 planes, we use reciprocal lattice vectors such that (1,0) and (0,1) are along the lines joining nearest-neighbour copper atoms. In this notation, the magnetic excitations are strongest near the $\mathbf{Q} = (1/2, 1/2)$ position. The most prominent feature of the magnetic excitations in the superconducting state of the $\text{YBa}_2\text{Cu}_3\text{O}_{6+x}$ system, where x controls the hole-doping level, is the ‘resonance’ (refs 3–6). The resonance occurs at $(1/2, 1/2)$ and at energies of 41 meV for optimally doped $\text{YBa}_2\text{Cu}_3\text{O}_{6.93}$, where T_c is at the highest value, and at 34 meV for underdoped $\text{YBa}_2\text{Cu}_3\text{O}_{6.6}$, where x and T_c are below their optimal values^{3–6}. For energies below the resonance energy, E_{res} , the magnetic excitations are peaked away from $(1/2, 1/2)$ at the incommensurate positions $(1/2 \pm \delta, 1/2)$ and $(1/2, 1/2 \pm \delta)$, with $\delta = 0.105$ for $\text{YBa}_2\text{Cu}_3\text{O}_{6.6}$ (refs 12, 13).

Although the magnetic excitations associated with the ‘resonance’ are obviously important because they show dramatic enhancement below T_c , they contain only a small part of the total spectral weight^{5,6}. Thus we need to understand the higher-energy excitations better because they may mediate electron pairing. Experimentally, magnetic excitations have been observed for energies above E_{res} in underdoped $\text{YBa}_2\text{Cu}_3\text{O}_{6+x}$ for energies up to about 120 meV (refs 5, 14, 16–18). Neutron-scattering measurements, performed using a three-axis spectrometer at a reactor^{16,18}, have shown that for energies greater than the resonance energy $E > E_{\text{res}}$, the magnetic response also peaks away from the $(1/2, 1/2)$ position. However, technical limitations have meant that the wavevector dependence of the high-energy excitations has not been determined. Here we use a time-of-flight neutron spectrometer equipped with position-sensitive detectors to measure the $E > E_{\text{res}}$ excitations in well-characterized $\text{YBa}_2\text{Cu}_3\text{O}_{6.6}$ with $T_c = 62.7$ K and $E_{\text{res}} = 34$ meV (refs 5, 12, 13, 17).

Figure 1a–c shows images of the magnetic scattering for the energy $E = 85$ meV at $T = 10$ K, 66 K ($T_c + 3.3$ K) and 300 K. At high temperatures (Fig. 1a) the magnetic response consists of a single peak centred on the $(1/2, 1/2)$ position. As the temperature is lowered through T_c , a quartet of peaks is formed at the positions $\mathbf{Q}_\epsilon = (1/2 \pm \epsilon, 1/2 \pm \epsilon)$ and $(1/2 \pm \epsilon, 1/2 \mp \epsilon)$, with $\epsilon = 0.12 \pm 0.01$ reciprocal lattice units (r.l.u.) (for $E = 85$ meV). The pattern is very different from the low-energy scattering observed in the same $\text{YBa}_2\text{Cu}_3\text{O}_{6.6}$ sample in that the four peaks at high-energy $E \approx 85$ meV are rotated by 45 degrees with respect to those observed at lower energy $E \approx 24$ meV (refs 12–15). The low-energy incommensurate peaks are shown in Fig. 2f. Figure 1d–f shows cuts along the diagonals of Fig. 1a–c. Coherent high-energy excitations peaked at $(1/2, 1/2)$ are present at $T = 300$ K, the incommensurate structure appears between 300 and 66 K and is almost fully developed by $T = 66$ K = $T_c + 3.3$ K. For comparison, Fig. 1g–i shows Q – E maps of the intensity of the low-energy magnetic excitations measured for $\mathbf{Q} = (1/2, k)$. The trajectory of this path passes through the low-energy incommensurate peaks at $(1/2, 1/2 \pm \delta)$. As the temperature is lowered from 300 to 66 K ($T_c + 3.3$ K), the magnetic response becomes strong near $(1/2, 1/2)$ and $E_{\text{res}} = 34$ meV. Further lowering the temperature leads to the resonance becoming stronger and incommensurate peaks developing below E_{res} . In contrast, the high-energy incommensurate peaks are fully developed at T_c . Both the low-energy and the high-energy incommensurate peaks develop through a loss of spectral weight near $(1/2, 1/2)$ and at their respective energies as the temperature is lowered.

To obtain a more complete picture of the high-energy response we collected data at four energies above the resonance energy. The evolution of the magnetic response with energy is shown in Fig. 2a–f. For the lowest energy probed above the resonance, $E = 66$ meV, we observe a square structure with the vertices of the square pointing along the (110) and $(\bar{1}\bar{1}0)$ directions. There is evidence that the response is strongest near the edge of the square: a cut along $(1/2 - h, 1/2 + h)$ (see Fig. 2j) shows incommensurate peaks away from $(1/2, 1/2)$. The incommensurate peaks are most developed in the energy range 70–90 meV: Fig. 2b and c clearly shows four peaks. At the highest energy studied (Fig. 2a), $E = 105$ meV, the scattering still displays the square structure observed for $E = 66$ meV and is peaked away from $(1/2, 1/2)$. However, a four-peak structure is not discernible. Figure 2j–l shows cuts through the $(1/2, 1/2)$ position. The solid lines are fits to the phenomenological response function $\chi''(\mathbf{Q}, \omega) = \sum_{\mathbf{Q}_\epsilon} A_{\mathbf{Q}_\epsilon}^4 (\xi^2 + (\mathbf{Q} - \mathbf{Q}_\epsilon)^2)^{-2}$, where the sum is over the incommensurate wavevectors $\mathbf{Q}_\epsilon$. It is interesting to compare the wavevector and energy-integrated spectral weights of the high-energy response studied here with the resonance and low-energy incommensurate structure. We find $\langle m^2 \rangle = 0.12 \pm 0.02 \mu_B^2$ per formula unit (f.u.) for the resonance and sub-resonance structure ($24 < E < 44$ meV) and $\langle m^2 \rangle = 0.26 \pm 0.05 \mu_B^2$ f.u.^{−1} for the high-energy scattering pictured in Fig. 2a–d ($60 < E < 120$ meV). Thus

the higher-energy response has a significantly greater contribution to the total fluctuating moment than the resonance structure.

The present data show how structure develops in the high-energy magnetic excitation spectrum of an underdoped high- T_c copper oxide superconductor as it is cooled. The excitations develop their coherence between 300 and 66 K, that is, above the critical temperature T_c . Many other electronic properties, ranging from angle-resolved photoemission (ARPES)¹⁹ to resistivity, show corresponding changes above T_c for underdoped compositions. Such behaviour is generally attributed to a higher temperature scale T^* —the temperature at which the pseudogap develops²⁰.

Our data show that the high-energy magnetic excitations in the high-temperature superconductor $\text{YBa}_2\text{Cu}_3\text{O}_{6.6}$ consists of a

continuum of scattering bounded by a square and peaked at wavevector positions $(1/2 \pm \epsilon, 1/2 \pm \epsilon)$ and $(1/2 \pm \epsilon, 1/2 \mp \epsilon)$. A similar structure has recently been observed in the high-energy magnetic excitations of the magnetically ordered but weakly superconducting compound $\text{La}_{1.85}\text{Ba}_{0.125}\text{CuO}_4$ (ref. 10). This suggests there is universality, both in the low-energy^{12–15,21} and the high-energy spin dynamics between two very different classes of high- T_c superconductor. The magnetism in the $\text{YBa}_2\text{Cu}_3\text{O}_{6+x}$ system derives from the antiferromagnetic parent compound $\text{YBa}_2\text{Cu}_3\text{O}_6$. Thus it has been widely assumed²² that the high-energy dynamics are described by a short-range Heisenberg interaction with an exchange suitably reduced from the insulator value²³. This leads to the high-energy excitations being spin-wave-like, such as in the

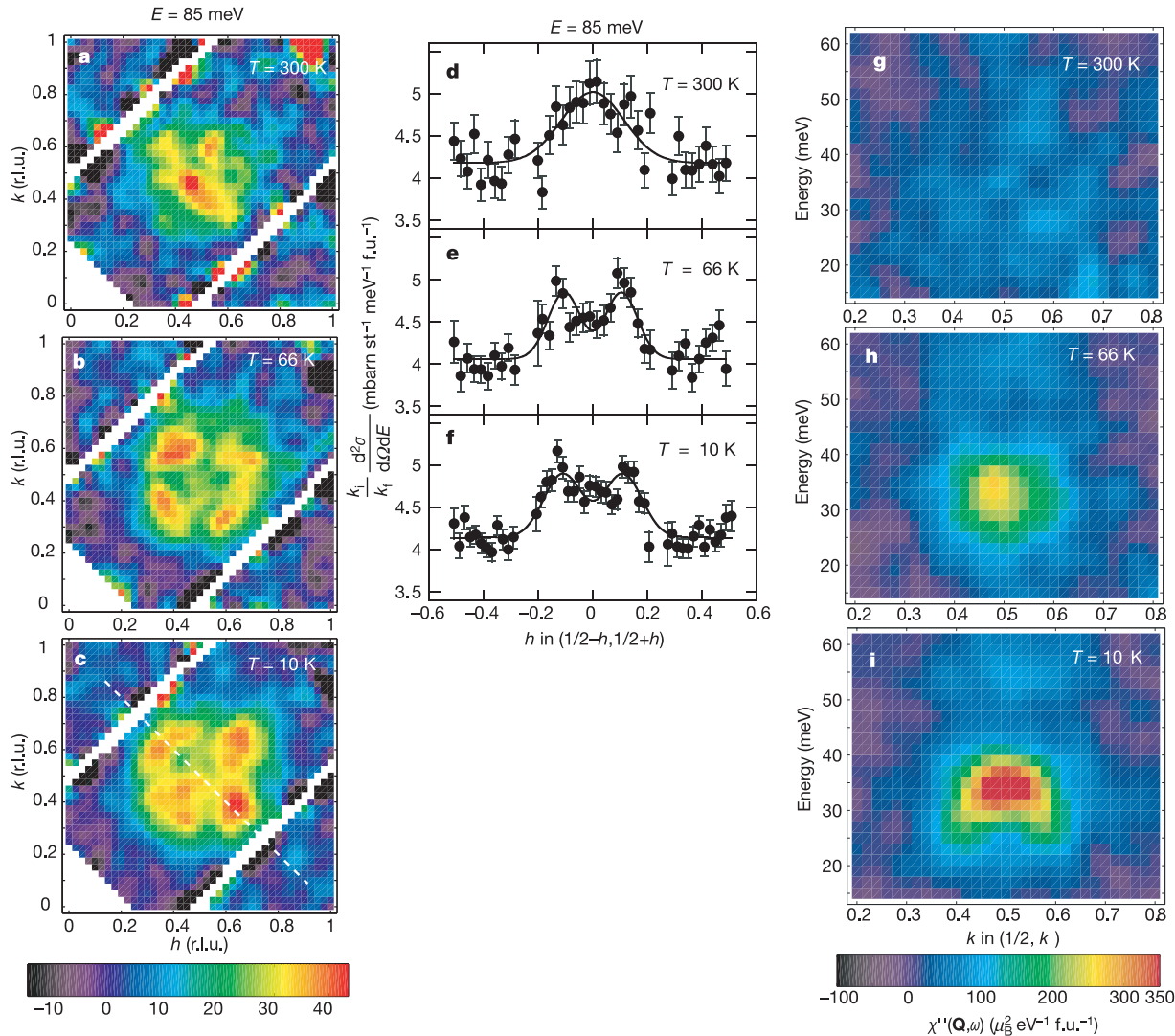


Figure 1 Images of the magnetic excitations in $\text{YBa}_2\text{Cu}_3\text{O}_{6.6}$, for various temperatures, measured by neutron scattering. **a–c**, Magnetic scattering at 300, 66 and 10 K showing the formation of the high-energy incommensurate peaks for $E = 85$ meV. **d–f**, Cuts through the incommensurate peaks along dashed line in **c** (no units conversion or background subtraction). A quadratic $(|\mathbf{Q}|^2)$ background has been subtracted and data converted to absolute units. The bilayer structure of $\text{YBa}_2\text{Cu}_3\text{O}_{6.6}$ means that the magnetic excitations depend on the out-of-plane wavevector. The data shown here correspond to acoustic or odd excitations and were measured with $l = 5.4$ r.l.u., where $\mathbf{Q} = h\mathbf{a}^* + k\mathbf{b}^* + l\mathbf{c}^*$, $\mathbf{a}^* = 2\pi/\mathbf{a}$, in which $\mathbf{b}^* = 2\pi/\mathbf{b}$, $\mathbf{c}^* = 2\pi/\mathbf{c}$. The lattice parameters (in Å) are $a = 3.82$, $b = 3.88$, and $c = 11.743$ for $\text{YBa}_2\text{Cu}_3\text{O}_{6.6}$. **g–i**, The magnetic response $\chi''(\mathbf{Q}, \omega)$ plotted as a function of wavevector and energy, for wavevectors along the dashed line in Fig. 2e. The data shown in **g–i** correspond to

acoustic or odd excitations with $1.4 < l < 2.0$. A measured energy-dependent background determined in the region $0 < h < 0.25$ and $0.3 < k < 0.7$ has been subtracted from the data. Experiments were performed using the MAPS chopper spectrometer at the ISIS spallation neutron source. Data were measured with an incident neutron energy of 160 meV and have a counting time of 48 h with a source proton current of $170 \mu\text{A}$. The 25-g single crystal sample was mounted in a thin-walled aluminium can. The MAPS spectrometer has an obstruction-free evacuated flight path to reduce background caused by air scattering. Data were generally collected at small scattering angles to minimize $|\mathbf{Q}|$, because the phonon and multiphonon scattering increases rapidly with $|\mathbf{Q}|$, while the magnetic scattering decreases with $|\mathbf{Q}|$. Data were placed on an absolute scale using a vanadium standard and corrected for the anisotropic $\text{Cu}^{2+} d_{x^2-y^2}$ form factor.

parent antiferromagnets above their Néel temperature. If such a picture were correct, our images (Fig. 2a–d) of the high-energy magnetic excitations should show rings of scattering (where the spin-wave dispersion intersects the corresponding constant-energy

plane) that get bigger as the energy is increased. The instrumental resolution (see Fig. 2a–f) would be sufficient to resolve the ring structure (see Supplementary Information). In fact, there is relatively little dispersion in the data in the energy range

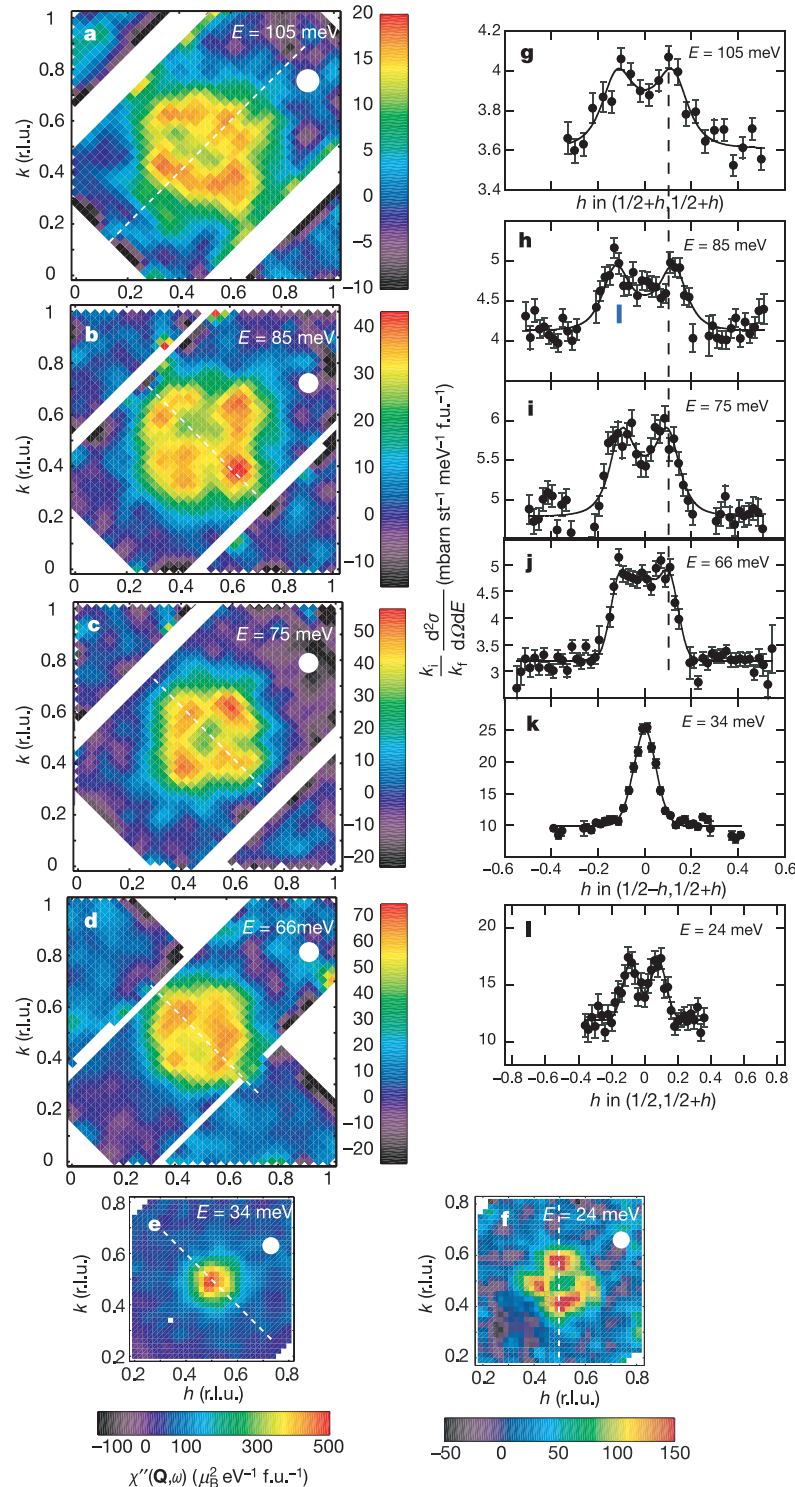


Figure 2 Images of the magnetic scattering in $\text{YBa}_2\text{Cu}_3\text{O}_{6.6}$ at $T = 10$ K for various excitation energies (frequencies). **a–f**, Magnetic excitations for energies between 24 and 105 meV. The data shown here correspond to acoustic or odd excitations and were measured with $l = 1.8$ or 5.4 r.l.u. **g–l**, Cuts through the $(1/2, 1/2)$ position along the dotted lines in **a–f**. The incident neutron energies used were 105, 160 and 220 meV. The $E = 105$ meV image was counted for 72 h. The white circles in **a–f** and the bar in **h** for

$E = 85$ meV represent the approximate instrumental resolution. The ranges of integration for energy transfers $E = 24, 34, 66, 75, 85$ and 105 meV are $\Delta E = 6, 6, 8, 10, 10$ and 30 meV, respectively. The solid lines in **g–j** are fits to the equation given in the text, where $\epsilon = 0.09(1), 0.10(1), 0.12(1), 0.11(1)$ r.l.u. and $\xi = 0.09(1), 0.09(1), 0.10(1)$ and $0.10(2)$ r.l.u. for $E = 66, 75, 85$ and 105 meV respectively. The vertical dashed line through **g–j** is the average value of ϵ .

$66 < E < 105$ meV (see dashed line in Fig. 2g–j). Thus a spin-wave-like dispersion at high energies seems to be inconsistent with our data.

Given that the observed spin dynamics is qualitatively similar in the two classes of high- T_c superconductor where they have been studied, one might expect that they have a common underlying origin. Candidates for this common origin include an incipient spin-charge separation leading to unidirectional stripes^{10,22} and the underlying electronic (band) structure. The high-energy magnetic excitations of stripes are strong along streaks in reciprocal space: these streaks can intersect to yield square-like structures. Alternatively, in an itinerant picture, the starting point for understanding the spin dynamics^{24,25} are electron–hole pair excitations about an underlying Fermi surface²⁶. This interpretation is supported by theoretical calculations performed using an itinerant approach based on a standard random phase approximation (RPA) that have predicted the correct geometry of the high-energy incommensurate peaks reported here²⁵.

The integrated spectral weight feature characterized here is considerably greater than the much sharper and widely discussed ‘resonance’ feature (Fig. 1j–i) present at lower energies^{3–6}. The high-energy excitations reported here are also present at T_c . Thus, in a magnetically mediated pairing mechanism the high-energy structure would have the dominant contribution to the pairing interaction. One might expect mediating excitations in the superconducting copper oxides to be a property of the CuO_2 planes themselves, and so they should be similar in different systems. The observation¹⁰ of high-energy magnetic excitations with a similar structure in $\text{La}_{1.875}\text{Ba}_{0.125}\text{CuO}_4$ supports this notion. $\square$

Received 29 October 2003; accepted 16 April 2004; doi:10.1038/nature02576.

1. Bardeen, J., Cooper, L. N. & Schrieffer, J. R. Theory of superconductivity. *Phys. Rev.* **108**, 1175–1204 (1957).
2. Chubukov, A. V., Pines, D. & Schmalian, J. in *The Physics of Superconductors* Vol. I, *Conventional and High- T_c Superconductors* (eds Binnemann, K. H. & Ketterson, J. B.) 495–590 (Springer, Berlin, 2003).
3. Rossat-Mignod, J. *et al.* Neutron scattering study of the $\text{YBa}_2\text{Cu}_3\text{O}_{6+x}$ system. *Physica C* **185**, 86–92 (1991).
4. Mook, H. A., Yethiraj, M., Aeppli, G., Mason, T. E. & Armstrong, T. Polarized neutron determination of the magnetic excitations in $\text{YBa}_2\text{Cu}_3\text{O}_7$. *Phys. Rev. Lett.* **70**, 3490–3493 (1993).
5. Dai, P. *et al.* The magnetic excitation spectrum and thermodynamics of high- T_c superconductors. *Science* **284**, 1344–1347 (1999).
6. Stock, C. *et al.* Dynamic stripes and resonance in the superconducting and normal phases of $\text{YBa}_2\text{Cu}_3\text{O}_{6.5}$ ortho-II superconductor. *Phys. Rev. B* **69**, 014502 (2004).
7. Scalapino, D. J. & White, S. R. Superconducting condensation energy and an antiferromagnetic exchange-based pairing mechanism. *Phys. Rev. B* **58**, 8222–8224 (1998).
8. Demler, E. & Zhang, S. C. Quantitative test of a microscopic mechanism of high-temperature superconductivity. *Nature* **396**, 733–735 (1998).
9. Hwang, J., Timusk, T. & Gu, G. D. High-transition-temperature superconductivity in the absence of the magnetic-resonance mode. *Nature* **427**, 714–717 (2004).
10. Tranquada, J. M. *et al.* Quantum magnetic excitations from stripes in copper oxide superconductors. *Nature* **429**, 534–538 (2004).
11. Axe, J. D. & Shirane, G. Influence of the superconducting energy gap on phonon linewidths in Nb_3Sn . *Phys. Rev. Lett.* **30**, 214–216 (1973).
12. Dai, P., Mook, H. A. & Dogan, F. Incommensurate magnetic fluctuations in $\text{YBa}_2\text{Cu}_3\text{O}_{6.6}$. *Phys. Rev. Lett.* **80**, 1738–1741 (1998).
13. Mook, H. A. *et al.* Spin fluctuations in $\text{YBa}_2\text{Cu}_3\text{O}_{6.6}$. *Nature* **395**, 580–582 (1998).
14. Arai, M. *et al.* Incommensurate spin dynamics of underdoped superconductor $\text{YBa}_2\text{Cu}_3\text{O}_{6.7}$. *Phys. Rev. Lett.* **83**, 608–611 (1999).
15. Bourges, P. *et al.* The spin excitation spectrum in superconducting $\text{YBa}_2\text{Cu}_3\text{O}_{6.85}$. *Science* **288**, 1234–1237 (2000).
16. Bourges, P. *et al.* High-energy spin excitations in $\text{YBa}_2\text{Cu}_3\text{O}_{6.5}$. *Phys. Rev. B* **56**, R11439–R11442 (1997).
17. Hayden, S. M. *et al.* Absolute measurements of the high-frequency magnetic dynamics in high- T_c superconductors. *Physica B* **241**–243, 765–772 (1998).
18. Fong, H. F. *et al.* Spin susceptibility in underdoped $\text{YBa}_2\text{Cu}_3\text{O}_{6+x}$. *Phys. Rev. B* **61**, 14773–14786 (2000).
19. Damascelli, A., Hussain, Z. & Shen, Z.-X. Angle-resolved photoemission studies of the cuprate superconductors. *Rev. Mod. Phys.* **75**, 473–541 (2003).
20. Timusk, T. & Statt, B. The pseudogap in high-temperature superconductors: an experimental survey. *Prog. Phys.* **62**, 61–122 (1999).
21. Christensen, N. B. *et al.* Universal dispersive excitations in the high-temperature superconductors. Preprint at (<http://arxiv.org/abs/cond-mat/0403439>) (2004).
22. Kivelson, S. A. *et al.* How to detect fluctuating stripes in the high-temperature superconductors. *Rev. Mod. Phys.* **75**, 1201–1241 (2003).
23. Shamoto, S., Sato, M., Tranquada, J. M., Sternlieb, B. & Shirane, G. Neutron scattering study of antiferromagnetism in $\text{YBa}_2\text{Cu}_3\text{O}_{6.15}$. *Phys. Rev. B* **48**, 13817–13825 (1993).
24. Bulut, N., Hone, D., Scalapino, D. J. & Bickers, N. E. Random-phase approximation analysis of NMR and neutron-scattering experiments on layered cuprates. *Phys. Rev. Lett.* **64**, 2723–2726 (1990).

25. Norman, M. R. Relation of neutron incommensurability to electronic structure in high-temperature superconductors. *Phys. Rev. B* **61**, 14751–14758 (2000).
26. Hussey, N. E., Abdel-Jawad, M., Carrington, A., Mackenzie, A. P. & Balicas, L. A coherent three-dimensional Fermi surface in a high-transition-temperature superconductor. *Nature* **425**, 814–817 (2003).

Supplementary Information accompanies the paper on www.nature.com/nature.

Acknowledgements This work was supported in part by the UK EPSRC, the US NSF and the DOE.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to S.M.H. (S.Hayden@bristol.ac.uk).

Quantum magnetic excitations from stripes in copper oxide superconductors

J. M. Tranquada¹, H. Woo^{1,2}, T. G. Perring², H. Goka³, G. D. Gu¹, G. Xu¹, M. Fujita³ & K. Yamada³

¹Physics Department, Brookhaven National Laboratory, Upton, New York 11973, USA

²ISIS Facility, Rutherford Appleton Laboratory, Chilton, Didcot, Oxon OX11 0QX, UK

³Institute for Materials Research, Tohoku University, Sendai 980-8577, Japan

In the copper oxide parent compounds of the high-transition-temperature superconductors¹ the valence electrons are localized—one per copper site—by strong intra-atomic Coulomb repulsion. A symptom of this localization is antiferromagnetism², where the spins of localized electrons alternate between up and down. Superconductivity appears when mobile ‘holes’ are doped into this insulating state, and it coexists with antiferromagnetic fluctuations³. In one approach to describing the coexistence, the holes are believed to self-organize into ‘stripes’ that alternate with antiferromagnetic (insulating) regions within copper oxide planes⁴, which would necessitate an unconventional mechanism of superconductivity⁵. There is an apparent problem with this picture, however: measurements of magnetic excitations in superconducting $\text{YBa}_2\text{Cu}_3\text{O}_{6+x}$ near optimum doping⁶ are incompatible with the naive expectations^{7,8} for a material with stripes. Here we report neutron scattering measurements on stripe-ordered $\text{La}_{1.875}\text{Ba}_{0.125}\text{CuO}_4$. We show that the measured excitations are, surprisingly, quite similar to those in $\text{YBa}_2\text{Cu}_3\text{O}_{6+x}$ (refs 9, 10) (that is, the predicted spectrum of magnetic excitations^{7,8} is wrong). We find instead that the observed spectrum can be understood within a stripe model by taking account of quantum excitations. Our results support the concept that stripe correlations are essential to high-transition-temperature superconductivity¹¹.

$\text{La}_{2-x}\text{Ba}_x\text{CuO}_4$ (‘Zurich’ oxide) is the material in which Bednorz and Müller¹ first discovered high-transition-temperature superconductivity. An anomalous suppression of the superconductivity found¹² in a very narrow region about $x = 1/8$ was later shown to be associated with charge and spin stripe order^{4,13}. Schematic diagrams of stripe order in the CuO_2 planes are shown in Fig. 1a and b. Between the charge stripes are regions with locally antiferromagnetic order. The ‘bond-centred’ stripe model shown has received considerable theoretical attention^{11,14–16}.

In a uniform (undoped), two-dimensional, antiferromagnetic CuO_2 plane, with lattice parameter $a = 3.8 \text{ \AA}$ between the nearest-neighbour Cu ions, a neutron diffraction measurement yields superlattice peaks characterized by the wave vector $\mathbf{Q}_{\text{AF}} = (1/2, 1/2)$, in units of $2\pi/a$. When vertical stripe order is present, the superlattice peaks move to $(1/2 \pm \delta, 1/2)$, where $\delta = 1/(2p)$, and p is the charge stripe spacing; $p = 4$ in the present case. Analogous peaks appear for the case of horizontal stripes (see Fig. 1f, g). In $\text{La}_{1.875}\text{Ba}_{0.125}\text{CuO}_4$, both stripe orientations are present, so we can observe both sets of superlattice peaks simultaneously (see Fig. 1i)¹³.

The ordered magnetic moments can fluctuate about their average orientations. It costs energy to create these fluctuations, and the variation of the magnetic excitation energy with wave vector $\mathbf{Q}$ can be measured by inelastic neutron scattering. Antiferromagnetic La_2CuO_4 exhibits conventional spin-wave behaviour^{17,18}, with an energy dispersion $\hbar\omega = c|\mathbf{Q} - \mathbf{Q}_{\text{AF}}|$ for $\hbar\omega < J$, with J ($\approx 140 \text{ meV}$)¹⁷ being the effective superexchange coupling between neighbouring magnetic moments. The spin-wave velocity c is proportional to J , and the spin waves extend up to a maximum

energy of $2J$. The strength of the magnetic scattering is relatively large throughout the antiferromagnetic Brillouin zone centred on $\mathbf{Q}_{\text{AF}}$ (indicated by the dashed-line diamond in Fig. 1f, g, i), but becomes negligible outside of that zone.

One might expect to find spin waves dispersing out of the superlattice peaks of $\text{La}_{1.875}\text{Ba}_{0.125}\text{CuO}_4$ (refs 7, 8), in the form of expanding cones as shown in Fig. 1d. The dispersion of spin waves has recently been measured in the stripe-ordered model compound $\text{La}_{2-x}\text{Sr}_x\text{NiO}_4$ (refs 19, 20), where $S = 1/2$ Cu spins are replaced by $S = 1$ Ni spins and the stripes run diagonally. In $\text{La}_{1.875}\text{Ba}_{0.125}\text{CuO}_4$ (ref. 13) and a related compound²¹, the low-energy excitations ($< 12 \text{ meV}$) behave similarly to spin waves. To find out what happens at higher energies, we have grown large, high-quality crystals of $\text{La}_{1.875}\text{Ba}_{0.125}\text{CuO}_4$ and studied them with a powerful time-of-flight neutron spectrometer, MAPS, at the ISIS spallation source.

Experimental measurements are presented in Fig. 2. These are constant-energy slices through the magnetic scattering, with the energy transfer, $E = \hbar\omega$, increasing from the lower left to upper right. The data are plotted within an antiferromagnetic Brillouin

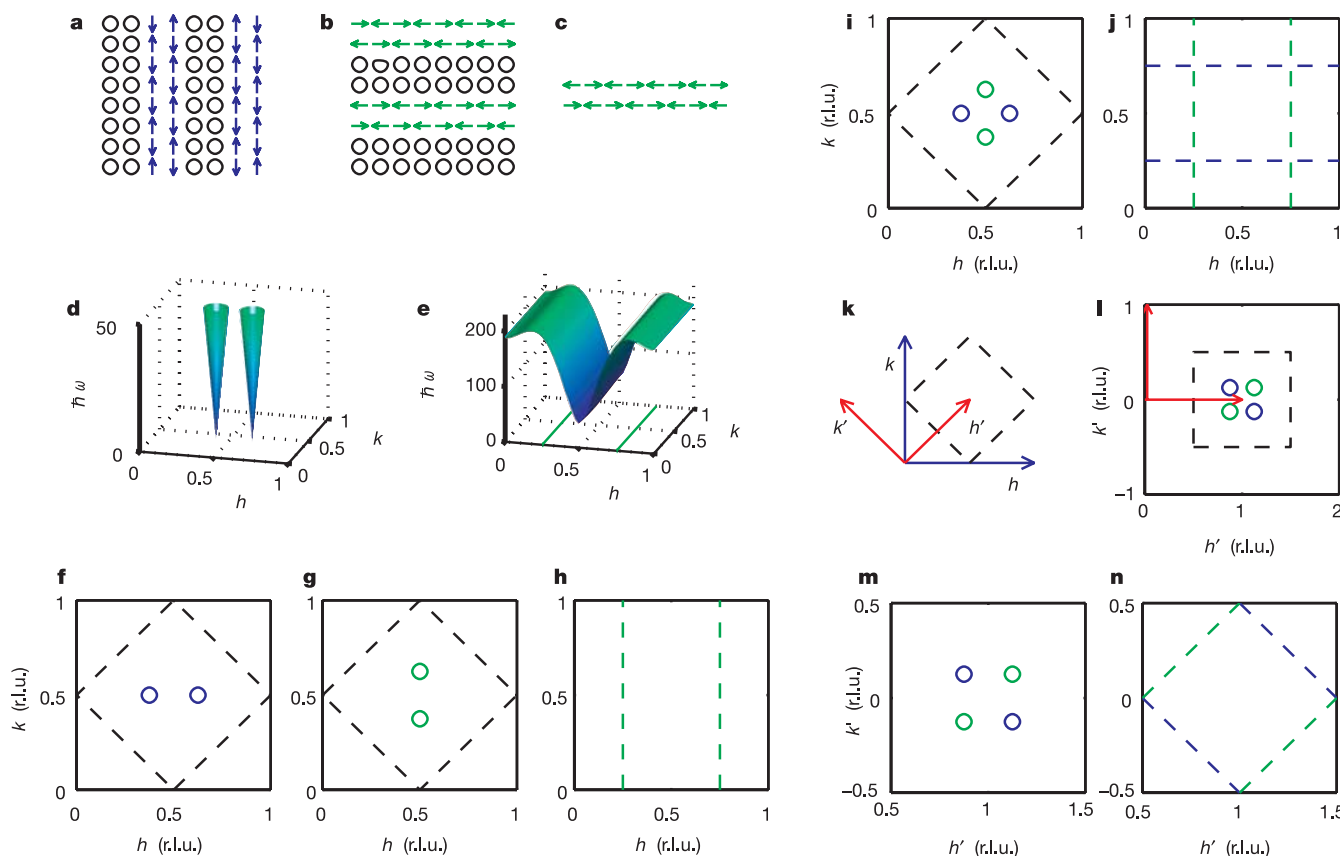


Figure 1 Schematic diagrams illustrating stripe order in real space and possible magnetic scattering in reciprocal space. **a**, Real-space diagram of vertical stripes; **b**, horizontal stripes; **c**, a horizontal spin ladder. Circles indicate Cu sites in hole-doped stripes, and arrows indicate magnetic moments on undoped Cu sites. (Although the specific alignment of the stripes with respect to the lattice has not been directly determined by experiment, the 'bond-centred' stripes shown here suggest a natural model for interpreting our results.) **d**, Dispersion of spin waves due to vertical stripes; **e**, triplet excitations in a horizontal spin ladder. For the triplet excitations, the minimum spin gap appears along $\mathbf{Q} = (1/2, k)$, for all k , and the maximum energy is reached along $\mathbf{Q}_m = (1/2 \pm 1/4, k)$ (ref. 25) as indicated by the lines in the $\hbar\omega = 0$ plane. Because of the singlet correlations, the magnetic scattering is relatively strong for $\mathbf{Q}$ between the lines $\mathbf{Q}_m$, but weak outside them. Thus, these lines define a zone of strong magnetic intensity. **f**, Reciprocal space,

with h and k in reciprocal lattice units (r.l.u.), showing incommensurate magnetic wave vectors for vertical stripe order (the dashed line gives the antiferromagnetic zone boundary); **g**, same for horizontal stripes; **h**, intensity zone boundary for a horizontal spin ladder. **i**, Wave vectors and boundaries expected when both orientations of stripes (and **j**, ladders) are present simultaneously. **k**, Rotation of the coordinate system by 45° changes the antiferromagnetic zone from a diamond into a square box, the latter being more convenient for plotting the data. **l**, Incommensurate points in the rotated system, where $\mathbf{Q}_{\text{AF}} = (1, 0, 0)$ measured in units of $2\pi/(a\sqrt{2})$. (We use the prime symbol ' to indicate wave vectors in the rotated system.) **m**, Expanded view of a single antiferromagnetic zone in the rotated system; **n**, same showing the zone of strong magnetic scattering for ladders, which forms a diamond.

zone, making use of the rotated coordinate system indicated at the bottom of Fig. 1. The magnetic scattering is centred on $\mathbf{Q}'_{AF}$ (where the prime ' denotes wave vectors in the rotated system). At 6 meV, we can clearly see the four incommensurate peaks, corresponding to fluctuations about the ordered stripe state. These low-energy results are similar to what is observed in superconducting $\text{La}_{2-x}\text{Sr}_x\text{CuO}_4$ (ref. 22). By 36 meV, the signal has dispersed inwards

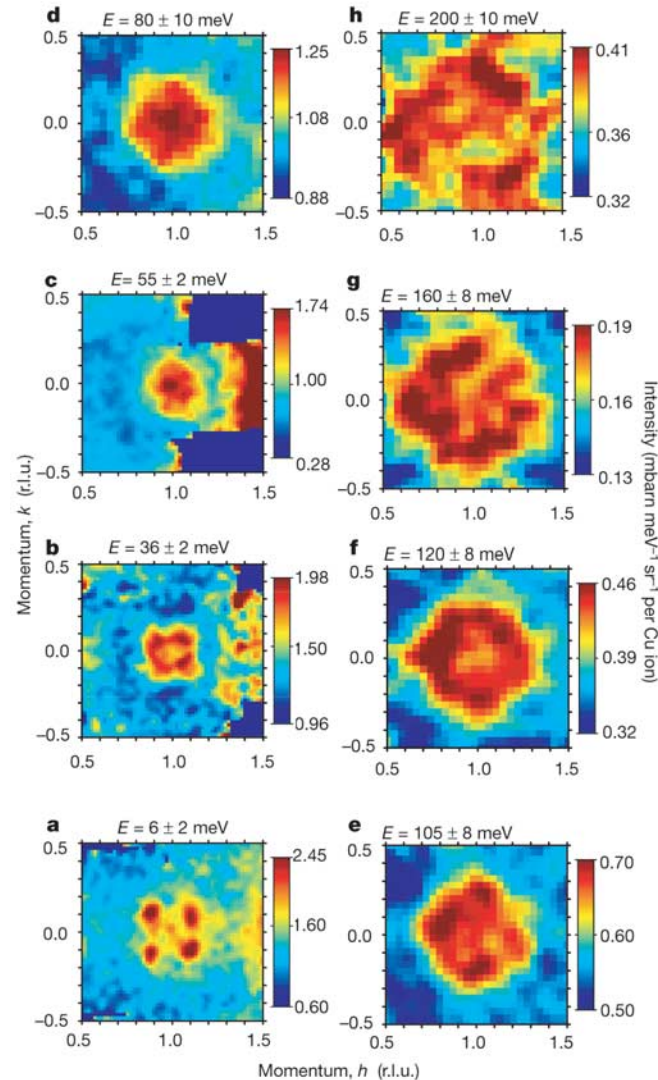


Figure 2 Constant-energy slices through the experimentally measured magnetic scattering from $\text{La}_{1.875}\text{Ba}_{0.125}\text{CuO}_4$. Intensity, measured at $T = 12\text{ K} (> T_c)$, is plotted in false colour within a single antiferromagnetic zone (compare Fig. 1m and n). Energy has been integrated over the ranges indicated by the error bars, and $\mathbf{Q}$ dependence has been convolved with a gaussian function to reduce scatter. Panels **a–c** were measured with an incident neutron energy $E_i = 80\text{ meV}$, panels **d–g** with $E_i = 240\text{ meV}$, and panel **h** with $E_i = 500\text{ meV}$. (In addition to the magnetic scattering, there are also features due to phonons, and ‘background’ from single- and multiple-phonon scattering.) The sample consisted of four crystals, with a total mass of 58 g, grown by the travelling-solvent floating-zone method at Brookhaven. Magnetic susceptibility measurements on pieces cut from the ends of each crystal show that T_c is generally less than 3 K, but at one end of each of two crystals it is close to 6 K. The crystal quality was first checked on the BT-9 spectrometer at the National Institute of Standards and Technology (NIST) Center for Neutron Research in Gaithersburg, Maryland, and the mounting and alignment were performed on the KSD spectrometer at the JRR-3M reactor in Tokai, Japan. For the experiment at the MAPS spectrometer, the co-aligned crystals were oriented with their c axes parallel to the incident beam.

towards $\mathbf{Q}'_{AF}$, nearly forming a ring about that point. A simple commensurate peak is found at 55 meV.

Using a spin-wave model, we would expect the 36-meV data to look something like Fig. 3a: four rings centred on the incommensurate wave vectors, representing cuts through the spin-wave dispersion cones. This picture is quite different from the experimental results. The behaviour at higher energies is even more striking. At 105 meV, the excitations have started to disperse outwards, but note the new shape: it is roughly a diamond, with points rotated 45° from the incommensurate wave vectors. This diamond continues to grow at 120 and 160 meV. Overall, the evolution of the magnetic scattering with energy looks very similar to that measured by Hayden *et al.*⁹ in superconducting $\text{YBa}_2\text{Cu}_3\text{O}_{6.6}$ (see Fig. 2 in ref. 9).

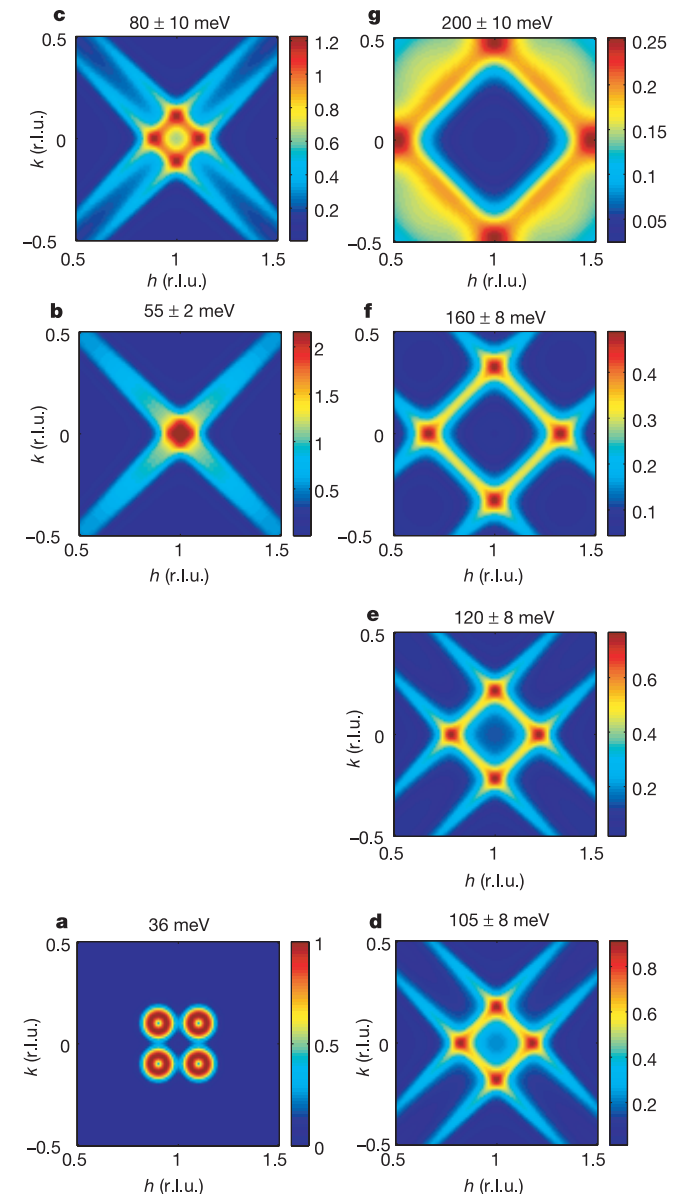


Figure 3 Simulations of constant-energy slices. **a**, Spin-wave model for $\hbar\omega = 36\text{ meV}$. **b–g**, Two-leg ladder model, discussed in the text, using $J = 100\text{ meV}$. For **b–g**, the δ -functions in frequency are replaced by lorentzians with energy width $\Gamma = 0.2J$. In all cases, we have averaged over both orientations (horizontal and vertical) of stripes or spin ladders, and have integrated over the same energy ranges as in Fig. 2.

The dispersion measured along $\mathbf{Q}' = (1 + q, q, 0)$ is presented in Fig. 4b. Another interesting quantity to consider is the function $S(\omega)$, obtained by integrating the magnetic scattering intensity $S(\mathbf{Q}, \omega)$ over $\mathbf{Q}$. The results are shown in Fig. 4a. With increasing energy, $S(\omega)$ initially decreases, and then rises to a broad peak near 50–60 meV. At higher energies, $S(\omega)$ gradually decreases. These results are qualitatively similar to earlier results on $\text{La}_{2-x}\text{Sr}_x\text{CuO}_4$ (ref. 23).

One generally determines the nature of magnetic fluctuations from the ordered state with which they are associated. In the case of La_2CuO_4 , the high-energy spin waves are clearly associated with the antiferromagnetic order. Although the excitations in our sample are clearly different from semiclassical spin waves, we nevertheless expect them to be associated with the stripe order indicated by magnetic and charge-order superlattice peaks¹³.

Is there a simple way to interpret our observations? If, for the moment, we ignore the low-energy incommensurate scattering, the finite-energy peak in $S(\omega)$ suggests that we are measuring singlet–triplet excitations of decoupled spin clusters. Given the stripe order in our sample, an obvious candidate for such a cluster would be one of the magnetic domains shown in Fig. 1a and b, corresponding to what is commonly called a two-leg spin ladder (Fig. 1c). (This name refers to the pattern formed by the exchange paths between the magnetic ions.) A spin ladder has the following interesting properties²⁴. The superexchange J between neighbouring spins keeps them antiparallel, but there is no static order at any temperature. This fluctuating, correlated state is said to be quantum disordered. There is a substantial energy gap ($\sim 0.5J$) to the first excited state, and the excitations disperse only along the ladder direction, not along the rungs (see Fig. 1e).

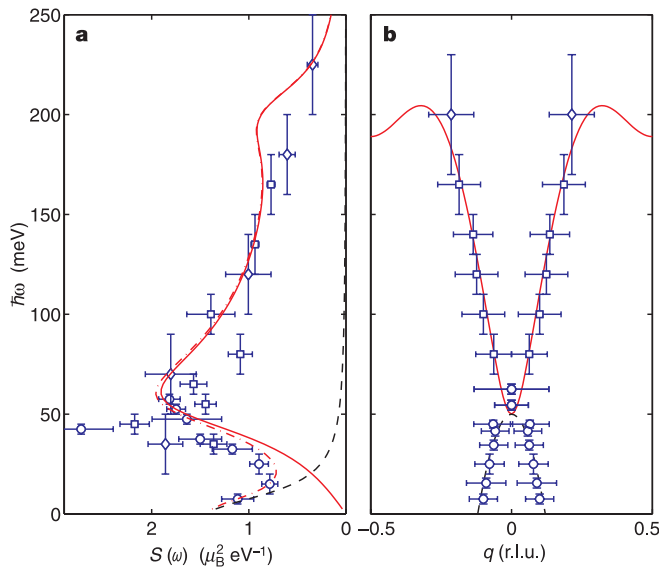


Figure 4 Experimental results for integrated magnetic scattering and dispersion of the excitations. **a**, $S(\omega)$, as defined in the text. Circles denote the $E_{\perp} = 80$ meV data set; squares denote $E_{\perp} = 240$ meV; diamonds denote $E_{\perp} = 500$ meV. In distinguishing the magnetic scattering from other signals, care was taken to avoid strong contributions from phonon branches at 20 and 47 meV. To obtain only the spin-dependent behaviour, we have corrected for the anisotropic magnetic form factor³⁰. Further investigation is required to determine whether or not the sharp feature at 42 meV is actually magnetic. **b**, Dispersion measured along $\mathbf{Q}' = (1 + q, q)$, with the assumption of symmetry about $q = 0$. Red lines in **a** and **b** are calculated from the two-leg spin ladder model with the same parameters as in Fig. 3. The black dashed line in **a** is a Lorentzian to describe the low energy signal, and the red dot-dashed line is the sum of the other two curves. Vertical 'error' bars in **a** and **b** indicate the energy range over which data were integrated, while horizontal bars in **b** indicate the half-widths in q of the fitted gaussian peaks.

To compare with experiment, we have calculated simulated spectra, see Fig. 3b–e, using the single-mode approximation for the scattering function of a spin ladder with isotropic exchange (I. Zalitznyak, unpublished work).

$$S(\mathbf{Q}, \omega) \approx (\hbar\omega_{q_{\parallel}})^{-1} [\sin^2(q_{\parallel}a/2) + \sin^2(q_{\perp}a/2)] \times [\delta(\omega - \omega_{q_{\parallel}}) - \delta(\omega + \omega_{q_{\parallel}})]$$

Here, $q_{\parallel}$ is measured parallel to the ladder, $q_{\perp}$ is along the rungs, and the dispersion $\omega_{q_{\parallel}}$, which is proportional to J , is given by ref. 25 (see Fig. 1e). Parts of this scattering function have been tested in measurements of ladder excitations on $\text{Sr}_{14}\text{Cu}_{24}\text{O}_{41}$ (ref. 26). In the simulations, we see that the most intense signal has a diamond shape that disperses outward with energy, similar to the right-hand side of Fig. 2.

The calculated and measured $S(\omega)$ and $\omega(q)$ are compared in Fig. 4a and b, respectively. The agreement is remarkable considering the simplicity of the model. The energy scale for the dispersion is set by a single parameter, J , and the value of J is only modestly reduced from that in the parent compound, La_2CuO_4 . The downward dispersion below 50 meV can be modelled by allowing weak coupling between the ladders, through the charge stripes, as demonstrated by the unpublished simulations of R. Konik and F. Essler.

For completeness, we note that dispersions with similarities to our data have also been obtained in weak-coupling, itinerant-electron calculations^{27,28}. Although this approach does provide a possible alternative explanation of our results, using it to explain the observed energy scale for the excitations would require fine tuning of parameters⁴, and the weak-coupling approach also does not explain the charge order in our sample. Thus, we believe that the ladder model, within the stripe picture, provides a more compelling explanation of the results. Given the similarity with recent measurements^{9,10} on $\text{YBa}_2\text{Cu}_3\text{O}_{6+x}$, together with the evidence for spatially-anisotropic magnetic excitations in detwinned $\text{YBa}_2\text{Cu}_3\text{O}_{6+x}$ (ref. 29), our results provide support for the concept that charge inhomogeneity, possibly dynamic in nature, is essential to achieve superconductivity with a high transition temperature in copper oxides^{5,11}. □

Received 26 January; accepted 16 April 2004; doi:10.1038/nature02574.

- Bednorz, J. G. & Müller, K. A. Possible high T_c superconductivity in the Ba-La-Cu-O system. *Z. Phys. B* **64**, 189–193 (1986).
- Anderson, P. W. New approach to the theory of superexchange interactions. *Phys. Rev.* **115**, 2–13 (1959).
- Orenstein, J. & Millis, A. J. Advances in the physics of high-temperature superconductivity. *Science* **288**, 468–474 (2000).
- Kivelson, S. A. *et al.* How to detect fluctuating stripes in the high-temperature superconductors. *Rev. Mod. Phys.* **75**, 1201–1241 (2003).
- Carlson, E. W., Emery, V. J., Kivelson, S. A. & Orgad, D. Concepts in high temperature superconductivity. In *The Physics of Superconductors Vol. II: Superconductivity in Nanostructures, High- T_c and Novel Superconductors, Organic Superconductors* (eds Bennemann, K. H. & Ketterson, J. B.) (Springer, Berlin, in the press); preprint at (<http://xxx.arxiv.org/pdf/cond-mat/0206217>) (2002).
- Bourges, P. *et al.* The spin excitation spectrum in superconducting $\text{YBa}_2\text{Cu}_3\text{O}_{6.85}$. *Science* **288**, 1234–1237 (2000).
- Kaneshita, E., Ichioka, M. & Machida, K. Spin and charge excitations in incommensurate spin density waves. *J. Phys. Soc. Jpn* **70**, 866–876 (2001).
- Batista, C. D., Ortiz, G. & Balatsky, A. V. Unified description of the resonance peak and incommensuration in high- T_c superconductors. *Phys. Rev. B* **64**, 172508 (2001).
- Hayden, S. M., Moock, H. A., Dai, P., Perring, T. G. & Dogan, F. The structure of the high-energy spin excitations in a high-transition-temperature superconductor. *Nature* **429**, 531–534 (2004).
- Reznik, D. *et al.* Dispersion of magnetic excitations in superconducting optimally doped $\text{YBa}_2\text{Cu}_3\text{O}_{6.95}$. Preprint at (<http://xxx.arxiv.org/pdf/cond-mat/0307591>) (2003).
- Arrigoni, E., Fradkin, E. & Kivelson, S. A. Mechanism of high temperature superconductivity in a striped Hubbard model. Preprint at (<http://xxx.arxiv.org/pdf/cond-mat/0309572>) (2003).
- Moodenbaugh, A. R., Xu, Y., Suenaga, M., Folkerts, T. J. & Shelton, R. N. Superconducting properties of $\text{La}_{2-x}\text{Ba}_x\text{CuO}_4$. *Phys. Rev. B* **38**, 4596–4600 (1988).
- Fujita, M., Goka, H., Yamada, K., Tranquada, J. M. & Regnault, L. P. Stripe order, depinning, and fluctuations in $\text{La}_{1.875}\text{Ba}_{0.125}\text{CuO}_4$ and $\text{La}_{1.875}\text{Sr}_{0.075}\text{CuO}_4$. Preprint at (<http://xxx.arxiv.org/pdf/cond-mat/0403396>) (2004).
- White, S. R. & Scalapino, D. J. Density matrix renormalization group study of the striped phase in the 2D t - J model. *Phys. Rev. Lett.* **80**, 1272–1275 (1998).
- Sachdev, S. & Read, N. Large N expansion for frustrated and doped quantum antiferromagnets. *Int. J. Mod. Phys. B* **5**, 219–249 (1991).

16. Zaanen, J., Osman, Y., Kruis, H. V., Nussinov, Z. & Tworzydło, J. The geometric order of stripes and Luttinger liquids. *Phil. Mag. B* **81**, 1485–1531 (2001).
17. Coldea, R. *et al.* Spin waves and electronic interactions in La_2CuO_4 . *Phys. Rev. Lett.* **86**, 5377–5380 (2001).
18. Kastner, M. A., Birgeneau, R. J., Shirane, G. & Endoh, Y. Magnetic, transport, and optical properties of monolayer copper oxides. *Rev. Mod. Phys.* **70**, 897–928 (1998).
19. Bourges, P., Sidis, Y., Braden, M., Nakajima, K. & Tranquada, J. M. High-energy spin dynamics in $\text{La}_{1.69}\text{Sr}_{0.31}\text{NiO}_4$. *Phys. Rev. Lett.* **90**, 147202 (2003).
20. Boothroyd, A. T. *et al.* Spin dynamics in stripe-ordered $\text{La}_{5/2}\text{Sr}_{1/2}\text{NiO}_4$. *Phys. Rev. B* **67**, 100407 (2003).
21. Ito, M. *et al.* Effects of “stripes” on the magnetic excitation spectra of $\text{La}_{1.48}\text{Nd}_{0.4}\text{Sr}_{0.12}\text{CuO}_4$. *J. Phys. Soc. Jpn* **72**, 1627–1630 (2003).
22. Cheong, S.-W. *et al.* Incommensurate magnetic fluctuations in $\text{La}_{2-x}\text{Sr}_x\text{CuO}_4$. *Phys. Rev. Lett.* **67**, 1791–1794 (1991).
23. Hayden, S. M. *et al.* Comparison of the high-frequency magnetic fluctuations in insulating and superconducting $\text{La}_{2-x}\text{Sr}_x\text{CuO}_4$. *Phys. Rev. Lett.* **76**, 1344–1347 (1996).
24. Dagotto, E. & Rice, T. M. Surprises on the way from one- to two-dimensional quantum magnets: the ladder materials. *Science* **271**, 618–623 (1996).
25. Barnes, T. & Riera, J. Susceptibility and excitation spectrum of $(\text{VO})_2\text{P}_2\text{O}_7$ in ladder and dimer-chain models. *Phys. Rev. B* **50**, 6817–6822 (1994).
26. Eccleston, R. S. *et al.* Spin dynamics of the spin-ladder dimer-chain material $\text{Sr}_{1.4}\text{Cu}_{2.6}\text{O}_{4.1}$. *Phys. Rev. Lett.* **81**, 1702–1705 (1998).
27. Kao, Y.-J., Si, Q. & Levin, K. Frequency evolution of neutron peaks below T_c : Commensurate and incommensurate structure in $\text{La}_{1.85}\text{Sr}_{0.15}\text{CuO}_4$ and $\text{YBa}_2\text{Cu}_3\text{O}_{6.6}$. *Phys. Rev. B* **61**, R11898–R11901 (2000).
28. Norman, M. R. The relation of neutron incommensurability to electronic structure in high temperature superconductors. *Phys. Rev. B* **61**, 14751–14758 (2000).
29. Mook, H. A., Dai, P., Dogan, F. & Hunt, R. D. One-dimensional nature of the magnetic fluctuations in $\text{YBa}_2\text{Cu}_3\text{O}_{6.6}$. *Nature* **404**, 729–731 (2000).
30. Shamoto, S., Sato, M., Tranquada, J. M., Sternlieb, B. J. & Shirane, G. Neutron-scattering study of antiferromagnetism in $\text{YBa}_2\text{Cu}_3\text{O}_{6.15}$. *Phys. Rev. B* **48**, 13817–13825 (1993).

Acknowledgements We thank A. R. Moodenbaugh and Q. Li for assistance with sample characterization, and acknowledge discussions with E. Carlson, F. Essler, S. A. Kivelson, R. Konik, S. Sachdev and I. Zalitznyak. J.M.T., H.W., G.D.G. and G.X. are supported by the Office of Science, US Department of Energy. K.Y. and M.F. are supported by the Japanese Ministry of Education, Culture, Sports, Science and Technology. Work was supported in part by the US–Japan Cooperative Research Program on Neutron Scattering.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to J.M.T. (jtran@bnl.gov).

A three-dimensional optical photonic crystal with designed point defects

Minghao Qi, Eleftherios Lidorikis, Peter T. Rakich, Steven G. Johnson, J. D. Joannopoulos, Erich P. Ippen & Henry I. Smith

Centre for Materials Science and Engineering, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, USA

Photonic crystals^{1–3} offer unprecedented opportunities for miniaturization and integration of optical devices. They also exhibit a variety of new physical phenomena, including suppression or enhancement of spontaneous emission, low-threshold lasing, and quantum information processing⁴. Various techniques for the fabrication of three-dimensional (3D) photonic crystals—such as silicon micromachining⁵, wafer fusion bonding⁶, holographic lithography⁷, self-assembly^{8,9}, angled-etching¹⁰, micromanipulation¹¹, glancing-angle deposition¹² and auto-cloning^{13,14}—have been proposed and demonstrated with different levels of success. However, a critical step towards the fabrication of functional 3D devices, that is, the incorporation of microcavities or waveguides in a controllable way, has not been achieved at optical wavelengths. Here we present the fabrication of 3D photonic crystals that are particularly suited for optical device integration using a lithographic layer-by-layer approach¹⁵. Point-defect microcavities are introduced during the fabrication

process and optical measurements show they have resonant signatures around telecommunications wavelengths (1.3–1.5 μm). Measurements of reflectance and transmittance at near-infrared are in good agreement with numerical simulations.

The photonic crystal (PhC) depicted in Fig. 1a is a layered structure that allows arbitrarily designed defects to be introduced in any layer. The layers are an alternating stack of the two complementary types of 2D-periodic photonic-crystal slabs: dielectric rods in air and air holes in dielectric. To our knowledge, the ‘woodpile’ structure (without intentional defects) is the only other such layered PhC successfully fabricated and measured at optical wavelengths^{5,6}. An advantage of the present structure is that the high symmetry of each layer enables future photonic devices to be realized by modifying only one layer. Moreover, defect states in this 3D PhC closely emulate their counterparts in 2D¹⁶. This allows a straightforward upgrade to 3D for most components and devices designed and analysed in 2D. Furthermore, the existence of two types of PhC slabs in this 3D crystal offers unprecedented polarization control¹⁷. Finally, another potential advantage is a large photonic bandgap (21% of the mid-gap frequency for circular holes¹⁵ and 25% for hexagonal holes, as shown in the Supplementary Fig. 1, versus 17% for the woodpile structure⁵).

Fabrication (Fig. 1b) takes a novel layer-by-layer approach where two layers (one hole layer and one rod layer, shown in the same colour in Fig. 1c) are fabricated in one process cycle, which consists of material deposition, aligned lithography, etching and planarization. Figure 1c illustrates the formation of the rod layer as a ‘by-product’ of over-etching the air cylinders (denoted ABCA) into the previously patterned layer. As a result, seven functional layers can be achieved in four process cycles. Details of the fabrication can be found in the Methods section. Highlights include the use of scanning-electron-beam lithography (SEBL) to align and define the patterns, and the use of spin-on-dielectric (SOD) to fill and planarize the etched cylinders. A set of dielectric point defects (unetched air-cylinders B’ in Fig. 1c), which can be viewed as microcavities in a 3D PhC, were introduced in the second process cycle. These defects have modes with low quality factors (Qs) of only about 25, but as such, are easily detectable in standard reflection and/or transmission measurements. In addition, many identical defects were introduced to further enhance the defect signatures. However, care must be taken that the defects should not either degrade the photonic bandgap or have strong interaction between them. A defect density of 15% in the plane was found to be reasonable. Furthermore, the lateral locations of these defects were non-periodic in order to suppress photonic sub-band structures and to yield simple signatures of the defect states. Frequency-domain calculations¹⁸ predict several cavity modes, and in particular three frequencies where resonant-transmission signatures are expected, around 1.3 μm , 1.4 μm and 1.5 μm , respectively. Despite the defect spanning three layers, the modes are all localized by the photonic bandgap near the middle of the structure. For example, a finite-difference time-domain (FDTD) calculation¹⁹ reveals the spatially localized electric field of a defect mode at 1.3 μm , as shown in Fig. 1d. The resulting fabricated structure is shown in Fig. 2. We note excellent layer-to-layer alignment (<40 nm) and the ability to achieve free-standing minimum feature size of 60 nm in a 3D nanostructure.

The PhCs were first characterized with a Nicolet Fourier-transform infrared (FTIR) microscope at room temperature. The unpolarized incident infrared light passes through a variable aperture, diverges to a hyperbolic reflecting mirror, and is then refocused to a small spot on the sample (Fig. 3a inset)²⁰. The incident light thus forms a cone-shaped bundle of rays with a half angle of 35°. The sample is located at the focal plane and can be tilted relative to the axis of the illumination cone. Reflectance and transmittance were taken for 0° sample tilt. Transmittances were also taken for samples tilted from [111] towards [11 $\bar{2}$] for 30° and 45°. Figure 3a shows

some typical transmission and reflection spectra. A stop-band is observed from $1.15\ \mu\text{m}$ to $1.6\ \mu\text{m}$ for all tilt angles. It is important to note that for non-zero sample tilt angles, each recorded spectrum is actually an average over a variety of incident angles. For example, a transmission spectrum at 45° sample tilt includes incident angles from 10° to 80° . Therefore these measurements clearly demonstrate an omni-directional photonic bandgap. Quantitatively, the reflectance and transmittance in the 0° sample tilt (corresponding to a small range of incident angles around 35° with respect to the sample normal) add up to $(100 \pm 6)\%$ for a wide wavelength range from $1.31\ \mu\text{m}$ up to $2.45\ \mu\text{m}$. Furthermore, the transmission and reflection curves cross at $1.68\ \mu\text{m}$ with exactly 50% transmission and 50% reflection. At short wavelengths, in the vicinity of the upper edge of the bandgap, reflectance drops down as theory predicts. The fact that an associated transmittance increase is not observed is most likely due to the fact that we were near the measurement noise floor of the spectrometer at those wavelengths. A stronger light source, as we shall describe shortly, indeed yielded high transmittance at short wavelengths. If we define the threshold as where transmittance is less than 25% and reflectance is more than 75%, the bandgap extends from $1.15\ \mu\text{m}$ to $1.45\ \mu\text{m}$, corresponding to 23% of the mid-gap frequency. This verifies the large 3D bandgap of our structure and is in good agreement with the theoretical prediction of 21% full bandgap¹⁵.

A pronounced defect signature can be clearly seen at $1.3\ \mu\text{m}$, in the middle of the 3D bandgap, as a dip in the reflection and corresponding increase in transmission (Fig. 3a). In a control experiment, a previously fabricated structure without defects did not show such a dip in the bandgap. Moreover, the independence of resonant frequencies on excitation angle is strong evidence for the microcavity states. To our knowledge, this is the first direct experimental observation of microcavity states in a 3D PhC at optical wavelengths.

To understand the defect states quantitatively, the measured feature sizes (using a scanning-electron microscope whose stage

motion is monitored by a laser interferometer) of a fabricated 3D PhC were fed into an FDTD simulator to calculate the reflection spectrum for a 35° incident angle. The simulation is detailed in the Methods section. A comparison of measured and calculated spectra is shown in Fig. 3b, with good agreement. Although this level of agreement has not been reported in any previous work, a more critical comparison reveals that the defect resonance at $1.3\ \mu\text{m}$ appears somewhat broader than expected. This can be accounted for by inhomogeneous broadening, as the measurements were averaged over many individual point-defects. Small structural variation of individual defects is unavoidable given the limitations of current nanofabrication. In addition, the finer signatures of the defect modes shown in the simulation, including those near $1.4\ \mu\text{m}$ and $1.5\ \mu\text{m}$, are not discernable in the measurement.

The above-mentioned issues were resolved with spatially localized measurements of transmission and reflection, which were performed with an ultra-bright, broadband ($1.1\ \mu\text{m}$ to $1.9\ \mu\text{m}$), unpolarized white light produced through supercontinuum generation²¹. A lensed optical-fibre was used to focus the light to a spot of $2.5\ \mu\text{m}$ diameter. For a defect density of 15%, a $2.5\text{-}\mu\text{m}$ spot would, on average, sample three defects. This significantly reduces the degree of inhomogeneous broadening for the defect resonances, allowing us to observe the spectral characteristics of individual defect states, which was not possible with the FTIR microscope.

A comparison of measurement and simulation for transmittance is shown in Fig. 4a. The resonance at $1.3\ \mu\text{m}$ demonstrates remarkable agreement between measurement and simulation in terms of resonance magnitude and line-width. The Q of this defect mode is found to be approximately 25 from simulation, which assumes no inhomogeneous broadening, whereas the measurement achieves a Q of approximately 18. Simulation also proves that this resonance peak does correspond to a localized defect state in the 3D PhC. The field pattern within the PhC, excited by a planewave at the resonance frequency of $1.3\ \mu\text{m}$, was calculated and is shown in Fig. 1d, which exhibits strong field localization at the defect. In addition, a new

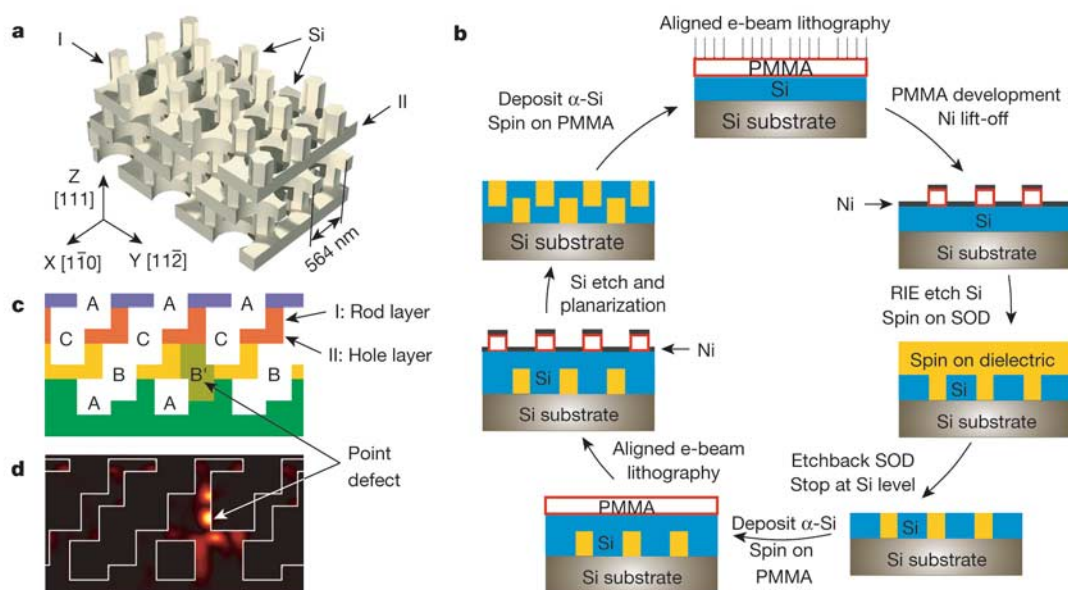


Figure 1 Schematic of the 3D PhC and its fabrication in silicon (Si). **a**, Computer rendering of an ideal 3D PhC with three hole layers and three rod layers. The fabricated structure has one additional hole layer on top of the uppermost rod layer. **b**, An illustration of the layer-by-layer approach, showing two process cycles and the over-etch into the previous layer. Details of the process flow can be found in the Methods section.

c, Cross-sectional view in the $[1\bar{1}0]$ plane. The PhC can be viewed as a face-centred-cubic lattice of air cylinders (denoted ABCA) in high dielectric (Si). Unetched air cylinders, denoted by B', form 'dielectric' point defects. **d**, Calculated $|\mathbf{E}|^2$ (where ϵ is the dielectric constant, and $\mathbf{E}$ is the electric field) of the point-defect state at a resonant wavelength of $1.3\ \mu\text{m}$. Note, localization of the electric field near the middle of the crystal region.

resonance peak near $1.5\ \mu\text{m}$, which was predicted by theory but not apparent in FTIR measurement, can now be identified in the measurement and is in good agreement with simulation.

The supercontinuum measurement at normal incidence is consistent with the FTIR measurement: it has similar attenuation in transmission but a wider bandgap, as expected from theory. Moreover, at short wavelengths, the transmittance does rebound. A comparison with the perfect 3D PhC without defects, whose transmission response was calculated and is shown as a dashed green line in Fig. 4a, is also noteworthy: the position and size of the bandgap does not change appreciably, indicating that the introduction of defects does not degrade the photonic bandgap.

A comparison of measurement and simulation for reflectance is shown in Fig. 4b. The FDTD simulation was performed for

an isolated defect using the supercell approximation (that is, a widely separated, very weakly interacting array of point defects). Remarkable agreement of the resonance frequencies is found in simulation and measurement, which are manifested as dips in the reflection spectrum at wavelengths around $1.3\ \mu\text{m}$, $1.4\ \mu\text{m}$ and $1.5\ \mu\text{m}$. In particular, the weak resonance at $1.4\ \mu\text{m}$ —which is predicted by theory but not identified in Figs 3 and 4a—is clearly seen and agrees very well with simulation. Quantitatively, a maximum reflectance of 85% is in reasonable agreement with the FTIR value of 90%. This lower reflectance, compared with both FTIR and simulation, is explained in part by difficulties in measuring absolute reflectivity due to the limited collection angle of the fibre lens, and also position sensitivity of the confocal arrangement.

All the agreements between simulations and measurements, for

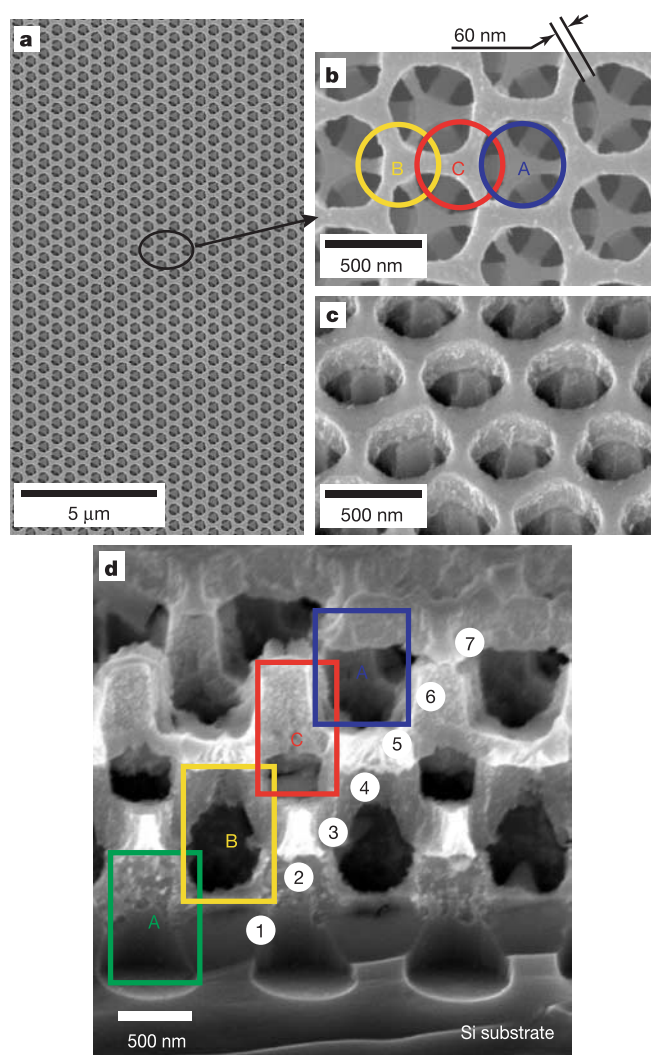


Figure 2 Scanning-electron micrographs of the fabricated 3D PhCs. **a**, Low-magnification top view. **b**, A close-up showing the hexagonal array of holes, denoted by A, and B and C are the sets of holes underneath the top layer, showing a sequential shift in the horizontal direction. The rod layers are completely blocked by the hole layers. **c**, A tilted top view reveals the rod layer. **d**, A cross-sectional view. The functional layers are marked by numbers, and the etched air cylinders are outlined with rectangles. The crystal shown here was fabricated in an early batch and exhibits higher surface roughness as well as more deviations from the designed parameters. It was not included for optical characterization.

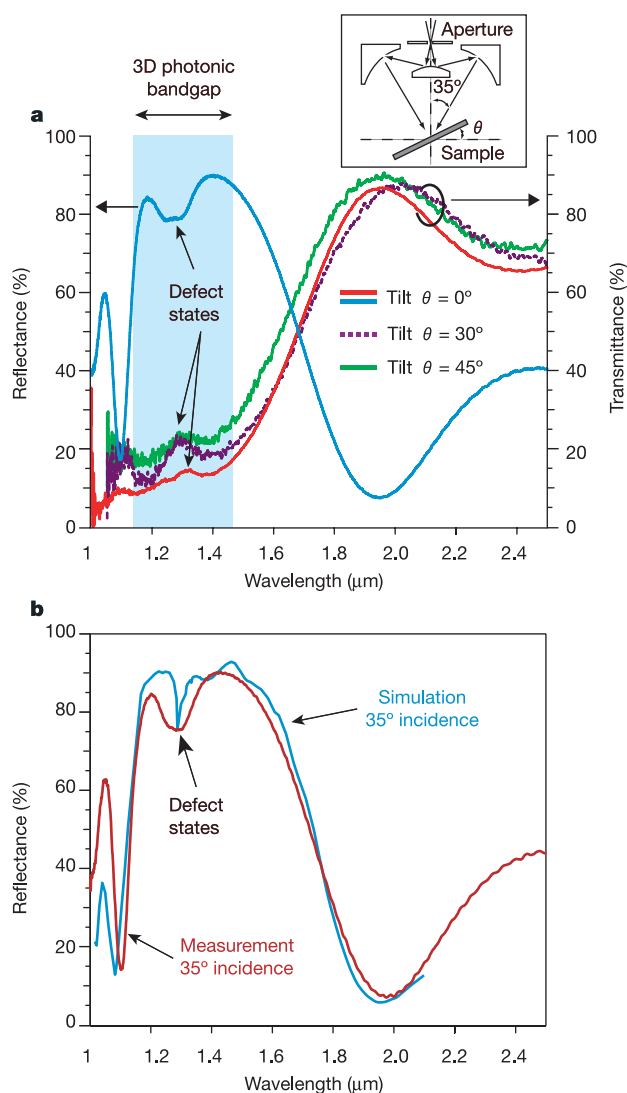


Figure 3 FTIR-microscopy characterization of the 3D PhC with point defects. **a**, Reflection (blue, at 0° sample tilt) and transmission (red, purple and green, for respective sample tilt angles of 0° , 30° and 45°). Inset is the optical path of the FTIR microscope. For non-zero degree sample tilts, the spectrum is an average of a variety of incident angles. **b**, Comparison of measurement (red) and simulation (blue) for reflection at 35° incident angle. No empirical adjustments of any kind were made other than the measured feature sizes. The wider linewidth of the measured defect state is attributed to inhomogeneous broadening.

both FTIR and the supercontinuum light source, are particularly noteworthy as no empirical adjustments of any kind (for example, fitting, shifting or rescaling of the spectra) were made either to theory or experiment. The only input into the simulation was the set of measured feature sizes.

Although the agreements of theory and experiment demonstrate the high quality of the fabricated PhC, fabrication imperfections such as surface roughness and misalignment were unavoidable, and indeed showed up in the characterization. At the upper part of the bandgap (wavelengths around 1.2 μm), the photonic crystal appears to produce some loss (that is, reflectance plus transmittance is less than one for both FTIR and supercontinuum measurements). This is expected, as at length scales comparable to the incident wavelength, scattering loss due to surface roughness becomes more significant. The size of the bandgap, when characterized with the supercontinuum source, showed narrowing at places where the

fabrication was known to be inferior, such as at the edge or corner of a die.

Finally, we note that the bandgap positions are quite sensitive to the dimensions of the fabricated structures. With the flexibility of electron-beam lithography, we investigated the tuning of bandgaps with structural variations. Specifically, 3D PhCs with the same lattice constant but different hole sizes, corresponding to different filling ratios, were fabricated and measured. The results confirmed the theoretical prediction that a lower Si-filling-ratio results in a shift of the bandgap towards shorter wavelengths.

One important discriminating feature of the lithographic approach used in this work is the ability to fabricate a 3D PhC with intentional defects in a natural, relatively low-cost and flexible manner. More generally, our fabrication process can be applied to a variety of 3D nanostructures. It is not material specific, and thus can be applied, in principle, to other material systems, and even possibly to epitaxial structures based on GaAs or InP. Furthermore, by using interference, X-ray or imprint lithography (refs 22–24, and C. Zanke, M. Qi and H. I. Smith, unpublished work), our layer-by-layer approach can produce large-area (several centimetres squared) and lower-cost 3D photonic crystals at near-infrared or even visible wavelengths. Defects can then be introduced at any level using SEBL.

The demonstration of designed defect states in a 3D photonic crystal, together with a robust fabrication process for general 3D nanostructures, should lead to more functional devices in the near future, and eventually to 3D optical integrated circuits. $\square$

Methods

Fabrication

Process flow (Fig. 1b): first, a layer of 520-nm amorphous silicon was deposited on a silicon wafer. An electron-beam resist, 230-nm-thick polymethylmethacrylate (PMMA), was then spun on. A 4-nm Ni layer was evaporated on top of the PMMA to avoid charging during the electron-beam registration and writing. The pattern data was compensated for proximity effects due to electron back-scattering. After the electron-beam lithography exposure and development, the image in PMMA was transferred to a 60-nm-thick Ni layer by lift-off. This layer was then used as a hard mask for reactive-ion etching (RIE) of silicon in a mixture of CF_4 and O_2 , carefully timed to achieve the desired depth. A type of spin-on-dielectric, hydrogen silsesquioxane (HSQ, Fox-16, Dow Corning), was then spun on to fill the holes and planarize the surface. The surface was carefully brought to the level of silicon by a combination of polishing and etchback. At this point, the sample was ready for another cycle of the identical fabrication sequence. After four such cycles, the spin-on-dielectric was etched away with a 10% hydrofluoric acid solution. All the processes were done at moderate temperatures ($<400^\circ\text{C}$) to minimize stress.

Scanning-electron-beam lithography: the tool was a custom system that combined parts from two IBM-donated systems²⁵. The write field of the SEBL system was 102.4 μm by 102.4 μm ; the photonic crystal occupied an area of roughly 70 μm by 70 μm . Alignment marks of 70-nm-thick Au were placed at the corners of the electron-beam write field to provide registration²⁵. Detection of these marks becomes more difficult as more Si is deposited over them, thus at each process cycle the alignment marks were regenerated. A total of seven aligned SEBL exposures were used to complete the four process cycles.

FTIR measurement

The aperture size of the FTIR microscope was fixed so that the sample collection area was around 30 μm by 30 μm . The signal-to-noise ratio was high, as the PhCs were fabricated on double-sided polished Si wafers. For the transmission measurements, the reference was an area of bare Si adjacent to the 3D PhC. Reference measurements were taken for each tilt angle, and the spectra were normalized accordingly. For the reflection measurements, the reference was a freshly evaporated 1- μm Au film on a Si wafer. To obtain quantitative measurements, the transmission signal was first maximized on the sample by focus adjustment, then the sample was moved laterally out of the beam path and the reference taken. The sample was then moved back into the beam path and the transmission spectrum recorded. Each measurement was integrated 512 to 1,024 times to increase the signal-to-noise ratio. For reflectance, signals were maximized for both the reference and the device by focus adjustment. The collected spectra were faithfully reproducible both for peak positions and absolute reflectance/transmittance.

Supercontinuum light measurement

Supercontinuum light, generated in a highly nonlinear fibre, is unpolarized with a spectrum range from 1.1 μm to 1.9 μm . Spatially localized transmission measurements at normal incidence were performed by placing the PhC at the focal plane of the lensed fibre (2.5- μm focus spot diameter). Light transmitted through the PhC was collected with a fibre optic placed at the back side of the wafer, which has a narrow collection angle of 5° (Fig. 4a, inset). This guarantees that transmission measurements only sample a small cone

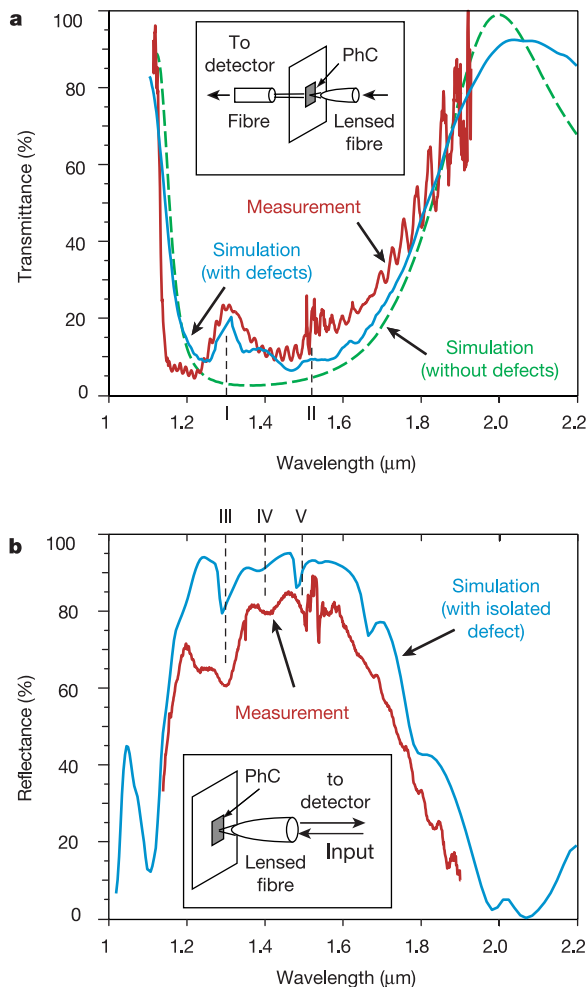


Figure 4 Supercontinuum-light-source characterization of point-defect states. The focal spot has a diameter of 2.5 μm , and samples on average three individual defects. The experimental setups are shown in the insets. **a**, Comparison between measurement (red) and simulation (blue) for transmission at normal incidence. Two defect resonance frequencies around 1.3 μm (I) and 1.5 μm (II), predicted by theory, can be clearly identified. The dashed green line is the calculated transmission with no defects. **b**, Comparison between measurement (red) and simulation for an isolated defect (blue) for reflection at normal incidence. Three resonant frequencies at 1.3 μm (III), 1.4 μm (IV) and 1.5 μm (V), can be clearly identified in both measurement and simulation.

of wave-vectors. The calibration reference was again an area of bare Si adjacent to the 3D PhC device. Spatially localized reflection measurements were performed by placing the lensed fibre at normal-incidence in a confocal geometry. The reflected light was collected and measured with the same fibre (Fig. 4b, inset).

Numerical simulation

We used the FDTD method to solve the full-vector time-dependent Maxwell's equations on a computational grid. This method is exact up to the discretization error (which in our case is about 1–2%). The only input used was the sample's feature sizes. The lateral size of the computational cell consists of a 16×8 arrangement of the minimum orthogonal unit cell and corresponds to a $9 \mu\text{m} \times 7.8 \mu\text{m}$ area. In the vertical direction, the cell consists of the seven PhC layers on top of a semi-infinite substrate. For amorphous Si in the range of $1.1 \mu\text{m}$ to $1.6 \mu\text{m}$ we used an average dielectric constant of 12. Defects were introduced in the second layer (see Fig. 1b) in a non-periodic distribution at a density of about 15%. Bloch-periodic boundary conditions were applied laterally and absorbing (perfectly matched layers) boundary regions vertically. A laterally plane and temporally gaussian source was introduced at the top of the computational cell, and fields and fluxes monitored both above and below the PhC. Upon Fourier transforming the fields and fluxes, we obtain the frequency-resolved reflectance and transmittance.

Received 20 December 2003; accepted 19 April 2004; doi:10.1038/nature02575.

1. Yablonovitch, E. Inhibited spontaneous emission in solid-state physics and electronics. *Phys. Rev. Lett.* **58**, 2059–2062 (1987).
2. John, S. Strong localization of photons in certain disordered dielectric superlattices. *Phys. Rev. Lett.* **58**, 2486–2489 (1987).
3. Joannopoulos, J. D., Meade, R. D. & Winn, J. N. *Photonic Crystals* (Princeton Press, Princeton, New Jersey, 1995).
4. Joannopoulos, J. D., Villeneuve, P. R. & Fan, S. Photonic crystals: putting a new twist on light. *Nature* **386**, 143–149 (1997).
5. Fleming, J. G. & Lin, S. Y. Three-dimensional photonic crystal with a stop band from 1.35 to 1.95 μm . *Opt. Lett.* **24**, 49–51 (1999).
6. Noda, S., Tomoda, K., Yamamoto, N. & Chutinan, A. Full three-dimensional photonic bandgap crystals at near-infrared wavelengths. *Science* **289**, 604–606 (2000).
7. Campbell, M., Sharp, D. N., Harrison, M. T., Denning, R. G. & Turberfield, A. J. Fabrication of photonic crystals for the visible spectrum by holographic lithography. *Nature* **404**, 53–56 (2000).
8. Vlasov, Y. A., Bo, X. Z., Sturm, J. C. & Norris, D. J. On-chip natural assembly of silicon photonic bandgap crystals. *Nature* **414**, 289–293 (2001).
9. Blanco, A. *et al.* Large-scale synthesis of a silicon photonic crystal with a complete three-dimensional bandgap near 1.5 micrometres. *Nature* **405**, 437–440 (2000).
10. Cheng, C. C. & Scherer, A. Fabrication of photonic band-gap crystals. *J. Vac. Sci. Technol. B* **13**, 2696–2700 (1995).
11. Aoki, K. *et al.* Three-dimensional photonic crystals for optical wavelengths assembled by micromanipulation. *Appl. Phys. Lett.* **81**, 3122–3124 (2002).
12. Kennedy, S. R., Brett, M. J., Toader, O. & John, S. Fabrication of tetragonal square spiral photonic crystals. *Nano Lett.* **2**, 59–62 (2002).
13. Kuramochi, E. *et al.* A new fabrication technique for photonic crystals: nanolithography combined with alternating-layer deposition. *Opt. Quant. Elec.* **34**, 53–61 (2002).
14. Sato, T. *et al.* Photonic crystals for the visible range fabricated by autocloning technique and their application. *Opt. Quant. Elec.* **34**, 63–70 (2002).
15. Johnson, S. G. & Joannopoulos, J. D. Three-dimensional periodic dielectric layered structure with omnidirectional photonic band gap. *Appl. Phys. Lett.* **77**, 3490–3492 (2000).
16. Povinelli, M. L., Johnson, S. G., Fan, S. & Joannopoulos, J. D. Emulation of two-dimensional photonic crystal defect modes in a photonic crystal with a three-dimensional photonic band gap. *Phys. Rev. B* **64**, 075313 (2001).
17. Lidorikis, E., Povinelli, M. L., Johnson, S. G. & Joannopoulos, J. D. Polarization-independent linear waveguides in 3D photonic crystals. *Phys. Rev. Lett.* **91**, 023902 (2003).
18. Johnson, S. G. & Joannopoulos, J. D. Block-iterative frequency-domain methods for Maxwell's equations in a planewave basis. *Opt. Express* **8**, 173–190 (2001).
19. Kunz, K. S. & Luebbers, R. J. *The Finite-Difference Time-Domain Methods* (CRC, Boca Raton, Florida, 1993).
20. Sting, D. W. & Messerschmidt, R. G. Reflective beam splitting objective. US patent 4,653,880 (1987).
21. Nicholson, J. W. *et al.* All-fiber, octave-spanning supercontinuum. *Opt. Lett.* **28**, 643–645 (2003).
22. Savas, T. A., Schattenburg, M. L., Carter, J. M. & Smith, H. I. Large-area achromatic interferometric lithography for 100 nm period gratings and grids. *J. Vac. Sci. Technol. B* **14**, 4167–4170 (1996).
23. Qi, M. & Smith, H. I. Achieving nanometer-scale, controllable pattern shifts in x-ray lithography using an assembly-tilting technique. *J. Vac. Sci. Technol. B* **20**, 2991–2994 (2002).
24. Chou, S. Y., Krauss, P. R. & Renstrom, P. J. Imprint lithography with 25-nanometer resolution. *Science* **272**, 85–87 (1996).
25. Boegli, V. & Kern, D. P. Automatic mark detection in electron beam nanolithography using digital image processing and correlation. *J. Vac. Sci. Technol. B* **8**, 1994–2001 (1990).

Supplementary Information accompanies the paper on www.nature.com/nature.

Acknowledgements We would like to thank M. Mondol and J. Daley for experimental assistance, and M. Povinelli for helpful discussions. The work was supported in part by a grant from the Materials Research Science and Engineering Center program of the National Science Foundation.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to M.Q. (mqi@alum.mit.edu).

Release of methane from a volcanic basin as a mechanism for initial Eocene global warming

Henrik Svensen¹, Sverre Planke^{1,2}, Anders Malthé-Sørensen¹, Bjørn Jamtveit¹, Reidun Myklebust³, Torfinn Rasmussen Eidem² & Sebastian S. Rey²

¹Physics of Geological Processes, University of Oslo, PO Box 1048 Blindern, 0316 Oslo, Norway

²Volcanic Basin Petroleum Research (VBPR), Oslo Research Park, 0349 Oslo, Norway

³TGS-NOPEC Geophysical Company, 3478 Nærnes, Norway

A 200,000-yr interval of extreme global warming marked the start of the Eocene epoch about 55 million years ago. Negative carbon- and oxygen-isotope excursions in marine and terrestrial sediments show that this event was linked to a massive and rapid ($\sim 10,000$ yr) input of isotopically depleted carbon^{1,2}. It has been suggested previously that extensive melting of gas hydrates buried in marine sediments may represent the carbon source^{3,4} and has caused the global climate change. Large-scale hydrate melting, however, requires a hitherto unknown triggering mechanism. Here we present evidence for the presence of thousands of hydrothermal vent complexes identified on seismic reflection profiles from the Vøring and Møre basins in the Norwegian Sea. We propose that intrusion of voluminous mantle-derived melts in carbon-rich sedimentary strata in the northeast Atlantic may have caused an explosive release of methane—transported to the ocean or atmosphere through the vent complexes—close to the Palaeocene/Eocene boundary. Similar volcanic and metamorphic processes may explain climate events associated with other large igneous provinces such as the Siberian Traps (~ 250 million years ago) and the Karoo Igneous Province (~ 183 million years ago).

A huge magmatic complex of dominantly subhorizontal sheets (sills) of basaltic composition intruded the Cretaceous Vøring and Møre basins before, and during, the northeast Atlantic continental break-up about 55 million years (Myr) ago^{5–7}. The sill complex covers an area of at least $80,000 \text{ km}^2$ (Fig. 1), but sills are also likely to be present in basin segments covered by lava flows that inhibit deep imaging (for example, beneath the $13,000 \text{ km}^2$ large 'Inner Flows' region). The thickness and number of sills is difficult to map as subsill seismic imaging is often poor. However, both field and seismic data commonly reveal several levels of sill intrusions^{6,8,9}. Three sills were drilled by the well 6607/5–2 with thicknesses of 2, 91 and $>50 \text{ m}$ (refs 6, 10). The middle sill is very well imaged on seismic reflection data, and can be followed for $>50 \text{ km}$ into the Vøring Basin. A conservative estimate of the volume of the sill complex in the Vøring and Møre basins is $0.9\text{--}2.5 \times 10^4 \text{ km}^3$ on the basis of an intruded area of $85,000 \text{ km}^2$ and an average vertical accumulated sill thickness of 100–300 m. The area of the entire sill complex in the North Atlantic Volcanic Province (NAVP) is probably at least five times greater than the sill complex in the Vøring and Møre basins. This estimate is based on the fact that the two studied basins comprise only 10% of the length of the northeast Atlantic volcanic margins, and a conservative estimate of the width of the NAVP sill complex being 50% of the width of the Vøring and Møre basins sill complex.

We have identified and characterized 735 hydrothermal vent complexes in the Vøring and Møre basins (Figs 1, 2 and 3). The total number of complexes in the Møre and Vøring basins is estimated to be at least 3–5 times greater, given the size distribution of the vent complexes and the seismic line coverage. This factor has

been confirmed by comparing the number of vent complexes identified on two- and three-dimensional (2D and 3D) seismic data in the same region. The hydrothermal vent complexes represent pipe-like structures with upper parts consisting of craters or mounds^{11–13}. Geochemical and petrographic data from the Vøring Basin and the Karoo Basin in South Africa show that vent complexes are affected by hydrothermal fluids (ref. 14 and Methods). Hydrothermal vent complexes are interpreted to originate in contact aureoles around sill intrusions, and formed by explosive release of fluids and sediments shortly (tens of years) after sill emplacement^{11–13}.

Constraints on the timing of magma emplacement, and thus vent complex formation, can be obtained from seismic interpretation and biostratigraphy. The seismic interpretation reveals that the intrusive volcanism occurred mainly before the main extrusive events because: (1) the Top Palaeocene horizon (the stratigraphic level where >95% of the hydrothermal vent complexes terminate) continues beneath the extrusive Inner Flows; and (2) no hydrothermal vent complexes have been identified within or above the extrusive sequence (Fig. 1). Careful interpretation shows that most vent complexes terminate at the same stratigraphic level, the Top Palaeocene, except for 20 vent complexes in the Møre Basin that terminate within the Palaeocene sequence.

New biostratigraphic dating of the hydrothermal vent complex drilled by 6607/12-1 (Fig. 3) shows that this vent complex was formed during the TP5a palynozone (55.0 to 55.8 Myr ago¹⁵), most probably at the start of the initial Eocene thermal maximum

(IETM). This interpretation is based on the presence of abundant immature *Apectodinium augustum* palynomorphs in a 25-m interval above the lower contact of the 'eye' structure and a few *in situ* mature *A. augustum* palynomorphs just below this interface. The increase of *A. augustum* is consistent with a worldwide increase of *A. augustum* during the IETM¹⁶, whereas the mature palynomorphs suggest a local heating event related to the formation of the hydrothermal vent complex that occurred just before the *A. augustum* bloom. Note that our timing is consistent with the timing of sill emplacement in the Faeroe–Shetland Basin obtained from seismic interpretation and biostratigraphic dating¹⁷.

Both field and seismic data suggest that the major part of the sill complexes were formed in a short time span. Individual sills in the NAVP may be up to hundreds of kilometres long and must have formed very rapidly (tens of years) to avoid solidification. Seismic observations indicate that a small number of large-volume intrusive episodes have formed the entire intrusive complex of the Vøring and Møre basins. For comparison, single flows with volumes exceeding 2,000 km³ have been estimated in the Columbia River Flood Basalt Province¹⁸ (that is, 4 to 12 such events may form the entire Vøring and Møre basins sill complex), but even larger individual events can be expected in the much more voluminous NAVP.

Our hypothesis of a rapid melt emplacement is further corroborated by volcanological arguments. The flux of melts into sedimentary basins is ultimately controlled by the rate of melt extraction from the mantle source. A conservative estimate of the vertical mantle melt flux is 0.005–0.1 m yr⁻¹. This figure is obtained by assuming that melt is produced by 5–10% partial melting, caused by

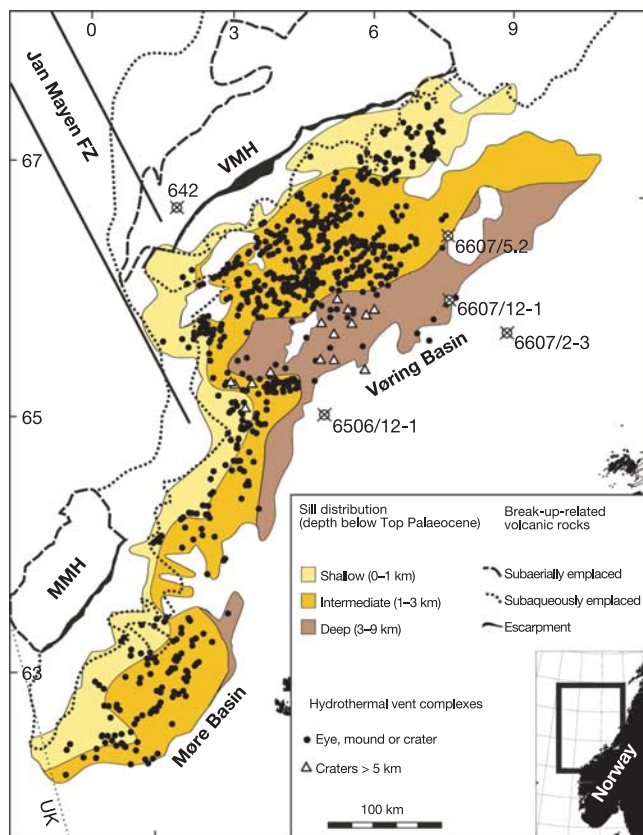


Figure 1 Distribution of hydrothermal vent complexes and volcanic intrusive and extrusive complexes on the mid-Norwegian continental margin. The map is based on detailed seismic mapping of >150,000 km of high-quality 2D seismic data and one 3D seismic survey (see Methods). Extrusive domains modified from ref. 31. The 'Inner Flows' is the subaqueous lavas emplaced landward of the escarpments. VMH, Vøring Marginal High; MMH, Møre Marginal High; FZ, fracture zone.

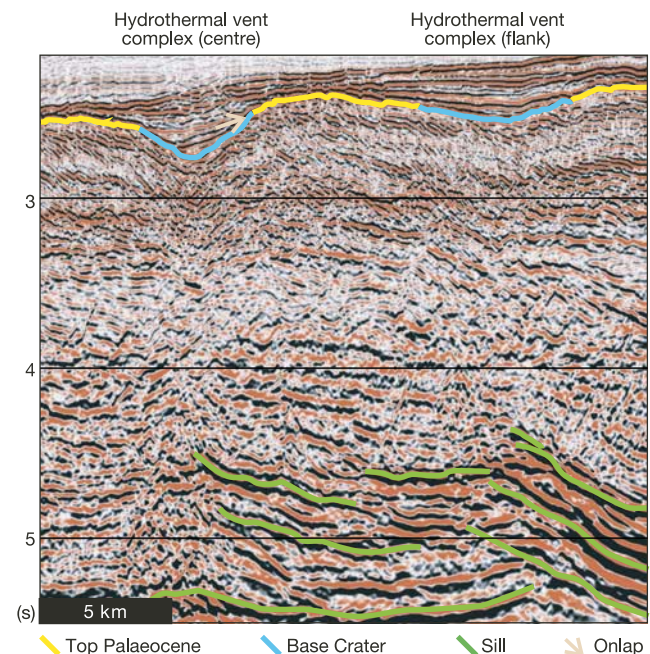


Figure 2 Seismic example of hydrothermal vent complexes in the southeast Vøring Basin. Two ~5 km-wide explosion craters are located at the Top Palaeocene level. The craters document the explosive power of the phreatic eruptions forming the hydrothermal vent complexes. Note the erosional base of the craters, onlapping sedimentary strata, disrupted chimney zones surrounded by inward-dipping strata below the craters, and a deep sill complex terminating below the craters. The craters are fairly rare (9%). The upper part of most hydrothermal vent complexes are eye-shaped (61%) or dome-shaped (30%), and less than 2 km in diameter (62%). Most hydrothermal vent complexes originate above the termination of a transgressive sill reflection, as shown by this example. Time shown in seconds (s) is the two-way travel-time for the seismic pulse.

decompression of mantle, that rises at a velocity of $0.1\text{--}1\text{ m yr}^{-1}$ (ref. 19). The calculated flux range is consistent with the flux obtained from observational data in the NAVP²⁰ (average melt flux of 0.0064 m yr^{-1} , assuming that two-thirds of the melt volume was emplaced in $500,000\text{ yr}$). The flux rates show that it will take $1,000\text{--}60,000\text{ yr}$ to produce the observed $100\text{--}300\text{ m}$ -thick sills complex without any focusing of mantle-derived melts.

The sills in the Møre and Vøring basins were mainly intruded into organic-rich Cretaceous and Palaeocene mudstones^{7,10}. The organic carbon is partly converted into methane when sedimentary rocks are heated beyond the gas window ($\sim 100\text{--}200^\circ\text{C}$; ref. 21). Accordingly, the methane production potential of an intruded sedimentary basin can be estimated from the total organic carbon (TOC) content of the sedimentary sequences and the volume and thermal history of the rocks heated by the intruding sills. We define a 'contact metamorphic aureole' as the volume of rocks heated beyond 100°C following the sill emplacement. As a rule of thumb, the thickness of the metamorphic aureole is comparable to, or greater, than the sill thickness on both sides of thick ($>50\text{ m}$) sills intruded into shales^{6,22,23}. The thermogenic methane produced from organic material in contact aureoles has a depleted carbon isotope ratio ($\delta^{13}\text{C} = -35\text{ to }-50\text{‰}$)²¹.

From Fig. 4 we estimate that $0.3\text{ to }3.0 \times 10^{18}\text{ g}$ of methane was produced in the metamorphic aureoles in the Vøring and Møre basins just after the sill emplacement. The methane production in the entire NAVP is possibly five times higher. It has been estimated that a release of $1.1 \times 10^{18}\text{ g}$ of CH_4 ($\delta^{13}\text{C} = -60\text{‰}$) is sufficient to cause the $>2.5\text{‰}$ negative excursion in carbon isotopes during the IETM³. Mass-balance calculations show that rapid release of the thermogenic methane produced in the metamorphic aureoles in the Vøring and Møre basins ($0.3\text{--}3.0 \times 10^{18}\text{ g CH}_4$) results in a carbon isotope excursion of $-0.2\text{ to }-3.0\text{‰}$ in the (present-day) exchangeable carbon reservoir (see ref. 3). Correspondingly, the

minimum excursion resulting from methane release in the entire NAVP is -1.0‰ .

Recent work indicates that surface temperatures across all latitudes rose by $6\text{ to }8^\circ\text{C}$ during the IETM, implying a rapid and massive release of carbon to the atmosphere²⁴. Extrusive volcanism accompanying and following the intrusive activity may have led to the release of vast amounts of greenhouse gases and aerosols in the NAVP (about 10^{20} g CO_2)²⁵. However, the extrusive phase took place over a $2\text{--}3\text{ Myr}$ period, leading to average fluxes of only $10^{13}\text{ g CO}_2\text{ yr}^{-1}$. Melt intruded into organic-rich sedimentary basins may cause considerably higher carbon fluxes into the atmosphere than a similar volume of erupted magma. Degassing of one cubic metre of CO_2 -saturated basaltic melt may release about 3.6 kg of carbon²⁶, whereas a melt intruded into organic-rich mudstones may trigger the release of $25\text{--}100\text{ kg}$ of carbon per cubic metre of magma. Hence, the carbon flux into the atmosphere would easily be more than an order of magnitude higher if the melt intruded into a relatively organic-rich sedimentary sequence rather than being erupted and degassed at the surface.

Contact metamorphism of organic material and associated methane venting in NAVP has the potential to cause the IETM $\delta^{13}\text{C}$ excursion if the methane is transported to the ocean or atmosphere. Our data suggest that the observed hydrothermal vent complexes represent efficient pathways for vertical fluid transport in sedimentary basins. In addition, gas from accumulations higher in the stratigraphy could be released when pierced by the

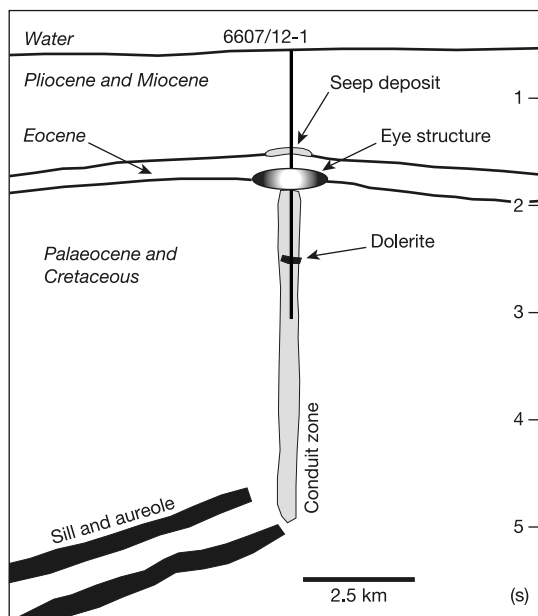


Figure 3 Schematic line drawing of a seismic profile across the hydrothermal vent complex drilled by 6607/12-1. A vertical zone of disturbed seismic data is interpreted as a conduit zone connecting the tip of a sill intrusion with the palaeo-surface. The eye-shaped structure represents the upper part of the complex, probably formed by filling of an explosion crater. Similar relationships between sills, conduit zone and palaeo-surface deposits are observed in most of the >700 hydrothermal vent complexes identified in the Vøring and Møre basins. Seep deposits above the eye shows that the vent complexes may be re-used for fluid migration a long time after their formation.

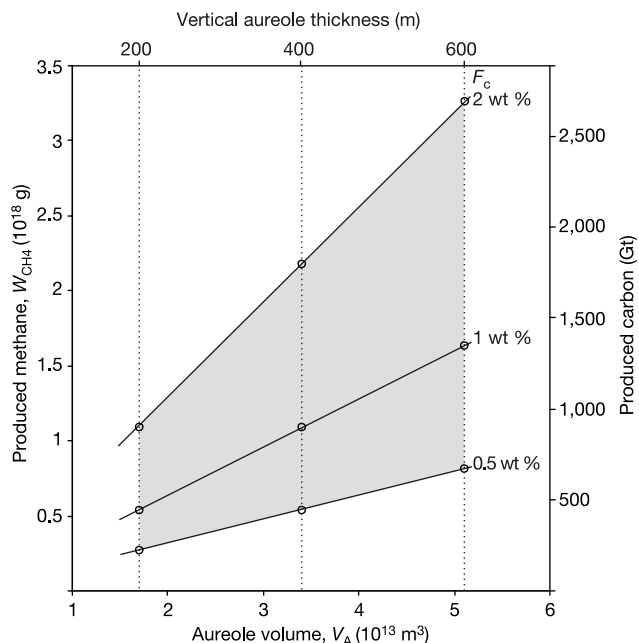


Figure 4 Estimated methane production potential of the Vøring and Møre basins. This is plotted as a function of the volume of the metamorphic aureole, V_A , and the amount of carbon transferred to methane, F_C . The mass of methane, W_{CH_4} , produced in a contact metamorphic aureole is:

$$W_{\text{CH}_4} = 1.34 F_C V_A \rho$$

where 1.34 is the atomic weight conversion factor between carbon and methane, and ρ is the rock density ($\rho = 2,400\text{ kg m}^{-3}$). The range of V_A is based on the mapped extent of the sill complex ($85,000\text{ km}^2$) and an average integrated vertical aureole thickness of $200\text{ to }600\text{ m}$. A realistic range of F_C is $0.5\text{ to }2.0\text{ wt \%}$ (see Methods). An estimate of the amount of gas ultimately produced from the original organic carbon during maturation and metamorphism is $50\text{--}90\%$, depending on the kerogen type²¹. Finally, note that we have not included the possible contribution from sediment decarbonation reactions, magma degassing, or shallow gas reservoirs pierced by the hydrothermal vent complexes.

hydrothermal vent complexes. Methane erupted beneath the sea from the hydrothermal vent complexes will probably reach the atmosphere. This is supported by numerical modelling showing that rising methane plumes formed by subaqueous gas eruptions provide efficient means of transporting methane through the water column without being dissolved or oxidized²⁷.

The seismic and biostratigraphic data strongly suggest a temporal correlation between the sill emplacement and the IETM, but the data do not have the ~10,000-yr resolution required to firmly establish this link. However, this could be proved by coring the Palaeocene–Eocene boundary strata on the flanks of one or more hydrothermal vent complexes, and identifying the negative carbon isotope excursion and its relation to deposits from the vent complexes.

We conclude that explosive release of metamorphic thermogenic methane during the intrusive phase of NAVP may have caused the extraordinary warming during the IETM. For reference, the average anthropogenic release of carbon during the 1990s was 6.3×10^{15} g carbon per year (ref. 28), which corresponds to a release of all the produced metamorphic CH₄ in the Vøring and Møre basins during a period of 35–360 yr.

Our observations stress the link between large igneous provinces (LIPs) and global climate changes. The climate impact of a LIP may be considerable if the melt is partly or completely emplaced into carbon-rich sedimentary successions. Volcanic basins thus provide a setting for rapid perturbations of the otherwise steady release of carbon from the sedimentary reservoirs. Several other major LIPs temporally correlated with prominent negative carbon isotope anomalies contain extensive subvolcanic intrusive complexes in carbon-rich sedimentary sequences, including the Siberian Traps (~250 Myr ago; the Permo-Triassic boundary), and the Karoo Igneous Province (~183 Myr ago; the Early–Middle Jurassic boundary) (compare with ref. 29). □

Methods

The seismic interpretation was done on a comprehensive seismic database consisting of more than 150,000 km of high-quality industrial multichannel seismic data and one 2,000 km³ 3D seismic survey using the Kingdom Suite seismic interpretation software. High-pass-filtered Bouguer gravity data and high-resolution aeromagnetic data were also loaded in the workstation project, and used to constrain the seismic interpretation.

The sill complex was carefully interpreted on the basis of seismic characteristics, in particular high reflection amplitude, saucer shape, local transgressive segments, and abrupt reflection termination. The interpretation was corroborated by analysis of potential field data and local pre-stack analysis of the seismic data, including re-processing and modelling. The sill complex interpretation was tied to well 6607/5-2, which penetrated three sill intrusions.

The hydrothermal vent complexes were similarly interpreted using seismic characteristics. The vent complexes are characterized by a circular eye-shaped, dome-shaped or crater-shaped upper part and a chimney-like lower part with disturbed seismic character commonly surrounded by inward-dipping reflections and bright reflections^{11–13}. Seismic evidence of fluid migration is commonly located in sedimentary strata above the vent complexes, such as chimneys, mounds, bright spots and flat spots. The disturbances sometimes continue all the way to the present sea floor. The location, size and seismic characteristics of each interpreted hydrothermal vent complex have been carefully described and stored in an electronic database. The seismic interpretation and mapping has been conducted in parallel with fieldwork on the intrusive and hydrothermal vent complexes in the Karoo Basin and numerical modelling studies of formation of hydrothermal vent complexes¹¹.

The well 6607/12-1 was drilled 2 km into the centre of a hydrothermal vent complex in 1986, and provides a unique opportunity to describe and analyse the properties of a vent complex. We have completed a comprehensive study of the seismic data, wireline log data, and core data and cuttings from this well. Samples (139 in total) were collected and analysed by scanning electron microscope and microprobe, X-ray diffraction (104 samples), isotope analysis (carbon, oxygen and strontium; 76 analyses), organic geochemistry (RockEval and total organic carbon; 67 samples), and vitrinite reflectance (78 samples). In addition, the biostratigraphy of the well was re-analysed, with particular focus on accurate dating of the Palaeocene and Eocene sequences. The data are partly described in ref. 12. Vitrinite reflectance data strongly suggest that the conduit zone of the hydrothermal vent complex is affected by hydrothermal fluids. The conduit zone is characterized by vitrinite reflectance values of up to 4.3%R_o, where %R_o is the vitrinite reflectance, compared with background values in the range 0.2–0.3%R_o.

Limited data on total organic carbon (TOC) values and kerogen type are published in the Vøring and Møre basins. TOC values in Upper Cretaceous sediments in wells 6607/2-3 and 6506/12-5 (Fig. 1) range from 0.4 to 2.6 wt % (mean = 1.4; standard

deviation = 0.44; N = 53)³⁰, but was probably higher during the Palaeocene. Shales with source-rock quality may also be present within the intruded sequences¹⁰.

Received 1 September 2003; accepted 4 April 2004; doi:10.1038/nature02566.

- Kennett, J. P. & Stott, L. D. Abrupt deep-sea warming, palaeoceanographic changes and benthic extinctions at the end of the Palaeocene. *Nature* **353**, 225–229 (1991).
- Schmitz, B. & Pujalte, V. Sea-level, humidity, and land-erosion records across the initial Eocene thermal maximum from a continental-marine transect in northern Spain. *Geology* **31**, 689–692 (2003).
- Dickens, G. R., O'Neil, J. R., Rea, D. K. & Owen, R. M. Dissociation of oceanic methane hydrate as a cause of the carbon isotope excursion at the end of the Paleocene. *Paleoceanography* **10**, 965–971 (1995).
- Dickens, G. R., Castillo, M. M. & Walker, J. C. G. A blast from the latest Paleocene: simulating first-order effects of massive dissociation of oceanic methane hydrate. *Geology* **25**, 259–262 (1997).
- Skogseid, J., Pedersen, T., Eldholm, O. & Larsen, B. T. Tectonism and magmatism during NE Atlantic continental break-up: the Vøring Margin. *Geol. Soc. Spec. Publ.* **68**, 305–320 (1992).
- Berndt, C., Skogly, O. P., Planke, S. & Eldholm, O. High-velocity breakup-related sills in the Vøring Basin, off Norway. *J. Geophys. Res.* **105**, 28443–28454 (2000).
- Brekke, H. The tectonic evolution of the Norwegian Sea Continental Margin with emphasis on the Vøring and Møre Basins. *Geol. Soc. Spec. Publ.* **167**, 327–378 (2000).
- Smallwood, J. R. & Maresh, J. The properties, morphology and distribution of igneous sills: modelling, borehole data and 3D seismic from the Faroe-Shetland area. *Geol. Soc. Spec. Publ.* **197**, 271–306 (2002).
- Chevallier, L. & Woodford, A. C. Morpho-tectonics and mechanism of emplacement of the dolerite rings and sills of the western Karoo, South Africa. *S. Afr. J. Geol.* **102**, 43–54 (1999).
- Brekke, H., Dahlgren, S., Nyland, B. & Magnus, C. The prospectivity of the Vøring and Møre basins on the Norwegian Sea continental margin. In *Petrol. Geol. NW Eur. Proc. 5th Conf.* (eds Fleet, A. J. & Boldy, S. A. R.) 261–274 (Geol. Soc. Publ., London, 1999).
- Jamveit, B., Svensen, H., Podladchikov, Y. & Planke, S. Hydrothermal vent complexes associated with sill intrusions in sedimentary basins. *Geol. Soc. Lond. Spec. Publ.* (in the press).
- Svensen, H., Jamveit, B. & Planke, S. Seep carbonate formation controlled by hydrothermal vent complexes: a case study from the Vøring volcanic basin, the Norwegian Sea. *Geo-Mar. Lett.* **23**, 351–358 (2003).
- Planke, S., Rasmussen, T., Rey, S. S. & Myklebust, R. Seismic characteristics and distribution of volcanic intrusions and hydrothermal vent complexes in the Vøring and Møre basins. In *Petrol. Geol. NW Europe and Global Perspectives—Proc. 6th Petrol. Geol. Conf.* (eds Tony Dore, T. & Vining, B.) (Geol. Soc. Publ., London, in the press).
- Woodford, A. C., et al. *Hydrogeology of the Main Karoo Basin: Current Knowledge and Research Needs* (Water Research Commission report 860, WRC, Pretoria, 2001).
- Berggren, W. A., Kent, D. V., Swisher, C. C. & Aubry, M.-P. A revised Cenozoic geochronology and chronostratigraphy. *Soc. Sed. Geol. Spec. Publ.* **54**, 129–212 (1995).
- Crouch, E. M. et al. Global dinoflagellate event associated with the late Paleocene thermal maximum. *Geology* **29**, 315–318 (2001).
- Trude, J., Cartwright, J., Davies, R. J. & Smallwood, J. New technique for dating igneous sills. *Geology* **31**, 813–816 (2003).
- Hopper, P. R. The Columbia River flood basalt province: Current Status. *AGU Geophys. Monogr.* **100**, 1–28 (1997).
- Weinberg, R. F. & Podladchikov, Y. Diapiric ascent of magmas through power-law crust and mantle. *J. Geophys. Res.* **99**, 9543–9559 (1994).
- Eldholm, O. & Grue, K. North Atlantic volcanic margins: dimensions and production rates. *J. Geophys. Res.* **99**, 2955–2968 (1995).
- Hunt, J. M. *Petroleum Geochemistry and Geology* (W. H. Freeman, San Francisco, 1996).
- Kjeldstad, A., Langtangen, H. P., Skogseid, J. & Bjørlykke, K. In *Advanced Topics in Computation Partial Differential Equations—Numerical Methods and Diffpack Programming* (eds Langtangen, H. P. & Tveito, A.) 611–656 (Springer, 2003).
- Galushkin, Y. I. Thermal effects of igneous intrusions on maturity of organic matter: a possible mechanism of intrusion. *Org. Geochem.* **26**, 645–658 (1997).
- Zachos, J. C. et al. A transient rise in tropical sea surface temperature during the Paleocene-Eocene thermal maximum. *Science* **302**, 1551–1554 (2003).
- Eldholm, O. & Thomas, E. Environmental impact of volcanic margin formation. *Earth Planet. Sci. Lett.* **117**, 319–329 (1993).
- Caldeira, K. & Rampino, M. R. The mid-Cretaceous super plume, carbon dioxide, and global warming. *Geophys. Res. Lett.* **18**, 987–990 (1991).
- Zhang, Y. Methane escape from gas hydrate systems in marine environment, and methane-driven oceanic eruptions. *Geophys. Res. Lett.* **30**, doi:10.1029/2002GL016658 (2003).
- IPCC (Intergovernmental Panel on Climate Change). *Climate Change 2001: The Scientific Basis* (IPCC, Geneva, 2001).
- Wignall, P. B. Large igneous provinces and mass extinctions. *Earth Sci. Rev.* **53**, 1–33 (2001).
- Eckhoff, E. E. *Foraminiferstratigrafi og Faunauvikling i Turbiditfacies fra Over Kritt på Midtnorsk Sokkel* Candidatus Scientiarum thesis, Dept Geol., Univ. Oslo (1999).
- Berndt, C., Planke, S., Alvestad, E., Tsikalas, F. & Rasmussen, R. Seismic volcanostratigraphy of the Norwegian Margin: constraints on tectonomagmatic break-up processes. *J. Geol. Soc. Lond.* **158**, 413–426 (2001).

Acknowledgements This study was supported by grants (H.S. and B.J.) from the Norwegian Research Council. We thank TGS-NOPEC for access to seismic data, H. Selnes for the biostratigraphic dating, and G. R. Dickens and J. G. Feder for comments on the manuscript.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to H.S. (hsvensen@geo.uio.no) or S.P. (planke@vbp.no).

Geochemical evidence from the Sudbury structure for crustal redistribution by large bolide impacts

James E. Mungall¹, Doreen E. Ames² & Jacob J. Hanley¹

¹Department of Geology, University of Toronto, 22 Russell St, Toronto, Ontario M5S 3B1, Canada

²Geological Survey of Canada, Natural Resources Canada, 601 Booth St, Ottawa, Ontario K1A 0E8, Canada

Deformation and melting of the crust during the formation of large impact craters must have been important during the Earth's early evolution, but such processes remain poorly understood¹. The 1.8-billion-year-old Sudbury structure² in Ontario, Canada, is greater than 200 km in diameter and preserves a complete impact section, including shocked basement rocks, an impact melt sheet and fallback material^{3,4}. It has generally been thought that the most voluminous impact melts represent the average composition of the continental crust⁴, but here we show that the melt sheet now preserved as the Sudbury Igneous Complex is derived predominantly from the lower crust. We therefore infer that the hypervelocity impact caused a partial inversion of the compositional layering of the continental crust. Using geochemical data, including platinum-group-element abundances, we also show that the matrix of the overlying clast-laden Onaping Formation represents a mixture of the original surficial sedimentary strata, shock-melted lower crust and the impactor itself.

The impact is inferred to have occurred in a shallow marine basin in the foreland of the Penokean Orogen because of the ubiquitous presence of clasts of carbonaceous (that is, rich in isotopically light organic carbon) mudstone⁵ in the upper kilometre of the fallback Onaping Formation. This possibly correlates with similar rocks of ~1,850 Myr ago from the southern Laurentian margin in Minnesota and Wisconsin^{6,7}, and the occurrence of impact-generated diamond⁸ in the Onaping Formation. The target crust beneath the mudstones comprised the Palaeoproterozoic Southern

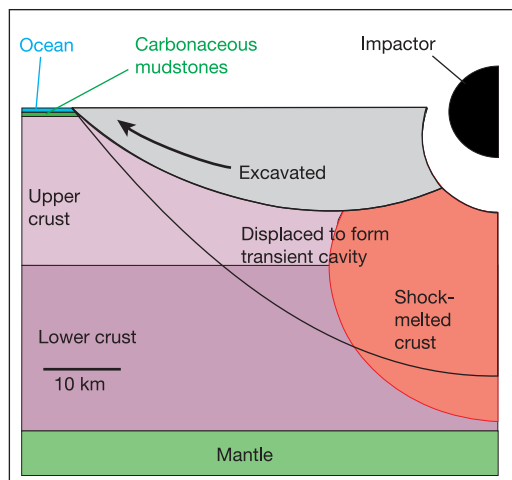


Figure 1 Schematic section of shock-melted and excavated volumes in ~40-km-thick continental crust after impact of an approximately 10-km bolide (after ref. 7). A large proportion of the shock-melted upper crustal region (shown in red) is within the volume of material that will be ejected immediately afterwards ('excavated' grey section), leaving the remaining impact melt dominated by lower crustal inputs.

Province fold belt and Archaean basement rocks of the Superior Province.

Passage of the shock wave generated by bolide impact caused nearly instantaneous melting of a roughly spherical volume of rock below the site of impact^{9,10} (Fig. 1). Much of the upper portion of the impact melt was ejected by the excavation, in the following few seconds, of a transient cavity that may have exceeded 30 km in depth⁴. The transient cavity then collapsed over a period of several minutes or hours to form a multi-ring basin 200 to 300 km in diameter, 1–6 km deep, wholly or partially occupied by a sheet of andesitic impact melt (the SIC), which differentiated internally into noritic and dioritic cumulates and a granitic residuum^{11,12}.

Brecciated, melted or vaporized target rocks and impactor were ejected to form a plume above the crater, before settling back onto the melt sheet to form the Onaping Formation¹³. The Onaping Formation includes, from bottom to top, discontinuous mass-flow deposits rich in arenitic metasedimentary clasts (<100 m, Garson member); suevitic fallback breccia with 60–90% blocky devitrified andesitic glass and lithic clasts dominated by igneous fragments (300 m, Sandcherry member); and a series of plume collapse units

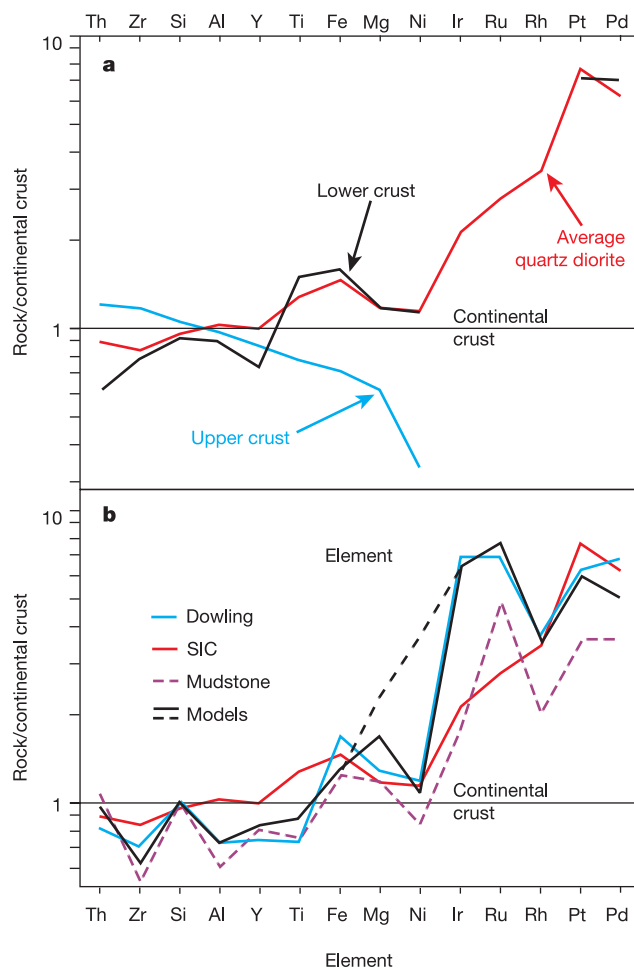


Figure 2 Compositions of the SIC and the Dowling member. **a**, Composition of the SIC (red) normalized to the average composition of bulk continental crust¹⁸, compared with the lower crust²³. Elements are shown in decreasing order of relative abundance in the upper crust¹⁸ (Th–Ni) or in conventional order of decreasing melting point (Ir–Pd). **b**, Average composition of four samples of the middle Dowling member compared with carbonaceous mudstones (average of four) and the SIC composition. Two models shown are mixtures of crustal material (70% carbonaceous mudstone, 30% SIC) with either 0.05% CI chondrite (solid black line) or 8% pyrolyte (dotted black line).

Table 1 Selected compositional data.

Description	<i>n</i>	Th (p.p.m.)	Zr (p.p.m.)	SiO ₂ (wt%)	Al ₂ O ₃ (wt%)	Y (p.p.m.)	TiO ₂ (wt%)	Fe ₂ O ₃ (wt%)	MgO (wt%)	Ni (p.p.m.)	Ir (p.p.b.)	Ru (p.p.b.)	Rh (p.p.b.)	Pt (p.p.b.)	Pd (p.p.b.)
Carbonaceous mudstone	4	9.0	110	62.10	9.10	20	0.50	7.67	6.88	47	0.09	0.49	0.12	1.44	1.45
Middle Dowling member	4	7.0	140	62.58	10.75	18	0.49	10.30	4.64	66	0.35	0.69	0.23	2.53	2.74
SIC quartz diorite	13	7.6	171	58.16	15.32	24	0.86	9.05	4.28	65	0.11	0.28	0.21	3.11	2.51

n, Number of samples included in average.

consisting of 30% lenticular devitrified glass bubble-wall shards and lithic clasts dominated by metasedimentary fragments (1,000 m, Dowling member), the upper 200 m of which is reworked owing to flooding by sea water. The Onaping Formation was emplaced containing about 0.4% native carbon, which is preserved in the Dowling member but has largely been oxidized by hydrothermal alteration in the Sandcherry member¹⁴.

Below the melt sheet, a sparse network was formed of dykes of shock-melted crustal rocks, which cooled to form the quartz diorite 'Offset' dykes^{15,16}, extending up to 50 km beyond the current outcrop area of the SIC. Offset dykes are coeval with the main mass of the melt sheet, and most were once present as melt-filled cracks extending beneath the originally horizontal melt sheet.

Later orogenies caused the burial, metamorphism, folding and exhumation of the Sudbury Structure to leave the SIC in its current configuration as a doubly plunging synformal remnant of the formerly horizontal melt sheet, transected by ductile shear zones with reverse and dextral displacements of several kilometres¹⁷.

Table 1 shows averaged partial compositions of carbonaceous metasedimentary clasts and samples of the matrix of the Dowling member of the Onaping Formation, and samples of quartz diorite from Offset dykes. Data were acquired by standard X-ray fluorescence methods on glass beads and pressed powder pellets. Platinum-group elements (PGE: Ir, Ru, Rh, Pt, Pd) were measured by NiS fire assay, acid digestion, and Te co-precipitation with inductively coupled plasma/mass spectroscopy (ICP-MS) finish at the Geoscience Laboratory of the Ontario Geological Survey in Sudbury. Detection limits are 0.03, 0.09, 0.03, 0.10, 0.06 and 0.27 p.p.b. for Ir, Ru, Rh, Pt, Pd and Au, respectively, for the 15-g samples used in this study¹⁸.

The main mass of the melt sheet was an essentially homogeneous sample of a large volume of the continental crust, showing no evidence for contributions from mantle-derived magmas^{9,16,19}. Offset dykes represent samples of the impact melt sheet, injected into the basement and chilled before significant differentiation had occurred¹⁶. The entire igneous stratigraphy of the differentiated SIC can be modelled as cumulates and residual liquids—all derived by closed-system crystallization of magma, like the Offset-dyke quartz diorites²⁰. In Fig. 2a we show the averaged concentrations of several relatively immobile major and trace elements in 13 samples of unmineralized inclusion-free quartz diorite from the Copper Cliff, Frood and Parkin Offset dykes. Similar major- and trace-element¹⁶ as well as Pt and Pd concentrations²¹ have been reported for unmineralized quartz diorites around the SIC, suggesting that these values are representative of the impact melt before separation of immiscible sulphide-ore liquid. Also shown for comparison are the average compositions of the upper and lower continental crust^{22,23}. Offset dykes are strongly enriched in transition metals compared with upper or average continental crust, but closely resemble the lower crust. This is understandable if the upper crustal portion of the shock-melted target rock volume was ejected by the crater excavation flow^{9,10}, leaving an impact melt sheet dominated by lower crustal inputs.

Our data suggest that in the currently exposed portion of the Sudbury Structure, the final result of the combination of relatively shallow excavation and deep-seated shock melting was the emplace-

ment of a sheet of lower crustal material several kilometres thick at the top of the crust, overlying what remained of the upper crust.

We also address the provenance of the devitrified glassy rocks of the Onaping Formation. Figure 2b illustrates the average compositions of clast-free samples of the matrix of the Dowling member of the Onaping Formation, selected from among several hundred as being the least affected by regional hydrothermal alteration^{13,14}. Shown for comparison is the average composition of four clasts of carbonaceous mudstone found as inclusions in the Onaping Formation and the SIC magma. The Dowling member is very similar to the average carbonaceous mudstone, and very different from either the average crust or the SIC, except for marked enrichment in all of the PGE. Note especially the exceptionally low concentration of Al₂O₃ in both the Dowling member and the carbonaceous mudstones (Table 1). The highly siderophile elements Ir and Ru are enriched in the Onaping Formation relative to any part of the continental crust, including the carbonaceous mudstone. This is shown again in Fig. 3—a plot of Ni versus Ir for samples of the

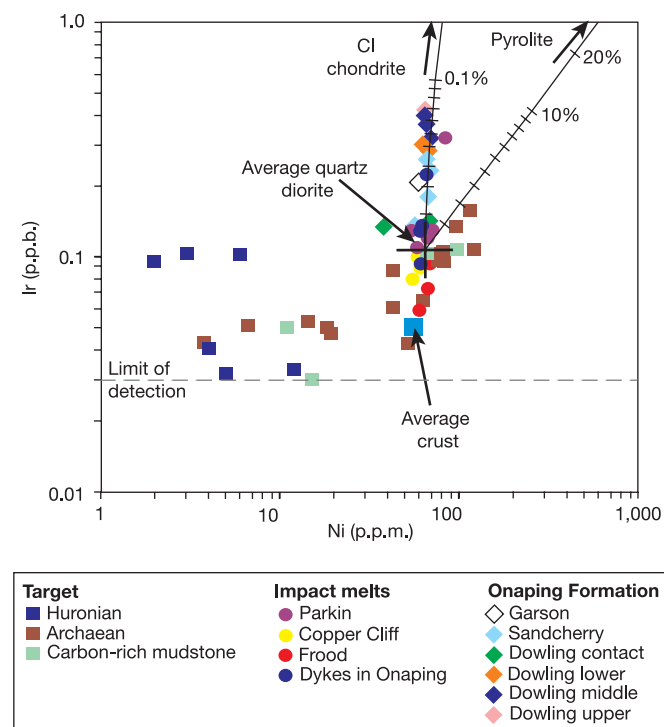


Figure 3 Ir and Ni contents of target rocks, impact melt, and Onaping Formation. The average quartz diorite (black cross) is near the upper limits of Ni and Ir contents of upper-crustal Proterozoic and Archean target rocks (metasediments, mafic, felsic igneous rocks). The Onaping Formation shows a general increase in Ir content with stratigraphic height, reaching up to 0.07% CI chondrite²⁶ in the upper Dowling member. Addition of sufficient pyrolyte²⁶ to the SIC to generate the Ir anomaly would cause excessive Ni enrichment. Note that much of the data for target rocks fell below the 0.03 p.p.b. detection limit for Ir.

Onaping Formation, quartz diorites and all major rock types in the Superior and Southern Province target crust. The intensity of the Ir contamination increases strongly with increasing height in the impact stratigraphy. The only two plausible sources for additional Ir are (1) the upper mantle of the Earth, which could have been included in the zone of shock melting if the impactor and crater were much bigger than currently supposed, or (2) the impactor itself, which has been proposed to have been a comet²⁴, and which should show chondritic ratios of non-volatile elements²⁵.

The best fit to the matrix of the middle Dowling member is obtained by taking a crustal component consisting of 30% quartz diorite as glass shards (lower crust) and 70% carbonaceous muddy matrix (upper crust) and adding to it either pyrolite or chondrite²⁶ in the indicated ratios. These proportions are in agreement with petrographic modal observations, given that we have excluded obvious lithic clasts from our samples. Both model compositions resemble the middle Dowling member for most elements; however, the use of pyrolite to account for the large positive Ir and Ru anomalies generates excessive collateral enrichment in Ni and Mg. The composition of the Dowling member therefore requires the presence of a meteoritic or cometary component and the absence of any detectable amount of shock-melted upper-mantle rocks.

The introduction of Ir and Ru in hydrothermal fluids cannot be ruled out, but hydrothermal transfer of PGE from igneous rocks to carbonaceous sediments conserves Pt/Ir ratios²⁷, whereas the enrichment of Ir in the Onaping Formation coincided with a drop in Pt/Ir of approximately a factor of ten relative to all possible crustal components. Osmium isotopic data would be of dubious value in the present context because of the ease with which Re–Os systematics can be reset during metamorphism⁶. If we were to dismiss the Ir enrichment observed in the impactite stratigraphy at Sudbury as being of hydrothermal origin, we must also question the meteoritic origins of the Ir enrichment in all impact-related lithologies, including those related to the Cretaceous–Tertiary boundary sequence^{28,29}.

The necessity of using shock-melted lower crust rather than the excavated upper crustal material in our mixing model for the matrix of the Dowling member demands an explanation, but current knowledge of cratering dynamics does not permit an unequivocal answer. We point out that (1) the streamlines in the crater-excavation flow field¹⁰ are directed away from the crater itself and hence upper crustal material should tend to leave the crater, and (2) hydrocode impact simulations show a kilometre-scale central rebound of impact melt several minutes after impact⁹. As the lower crustal rocks melted during impact should have contained approximately 1.5 wt% H₂O (ref. 23), we would expect the uprising low-pressure melt column to undergo explosive vesiculation and fragmentation, forming an ultraplinian eruption plume. Collapse of the plume would permit it to mix with sediment-charged sea water surging back into the crater. The Dowling Member thus contains a blend of glass bubble-wall shards with lower crustal provenance and redeposited muddy carbonaceous sediment from the uppermost crust^{5,30}.

Our work relates to the crustal-scale redistribution of matter during large impact events. Our first key observation is that the melt sheet originated in the lower crust, but was transported during collapse of the transient cavity to form a sheet several kilometres thick above what remained of the upper crust. The large-scale compositional zoning of the continental crust was therefore profoundly disturbed by the impact process. Intense Hadean meteorite bombardment might repeatedly have churned the lithosphere, hindering the stabilization of the crust until the rate of impact melting was much less than the rate of crustal differentiation through endogenic processes. Second, there is no indication that the impactor contributed to the composition of the impact melt sheet. Instead, the impactor was deposited as a veneer at the Earth's surface, mixed into the glassy ejecta and redeposited unconsolidated

sediments, its proportions increasing with stratigraphic height in the sequence that filled the crater. This is consistent with current models of accretion of a late veneer of chondritic material undepleted by interaction with core-forming metallic iron melts in the mantle. □

Received 29 October 2003; accepted 8 April 2004; doi:10.1038/nature02577.

- Melosh, H. J. & Ivanov, B. A. Impact crater collapse. *Annu. Rev. Earth Planet Sci.* **27**, 385–415 (1999).
- Krogh, T. E., Davis, D. W. & Corfu, F. in *The Geology and Ore Deposits of the Sudbury Structure* (eds Pye, E. G. *et al.*) 431–446 (Ontario Geological Survey, Special Vol. 1, 1984).
- Dietz, R. S. Sudbury Structure as an astrobleme. *J. Geol.* **72**, 412–434 (1964).
- Grieve, R. A. F., Stöffler, D. & Deutsch, A. The Sudbury structure: controversial or misunderstood? *J. Geophys. Res.* **96**, 22,753–22,764 (1991).
- Bunch, T. E. *et al.* in *Large Impacts and Planetary Evolution* (eds Dressler, B. O. & Sharpton, V. L.) 331–343 (Geological Society of America, Special Paper Vol. 339, 1999).
- Ripley, E. M., Park, Y.-R., Lambert, D. D. & Frick, L. R. Re-Os isotopic variations in carbonaceous pelites hosting the Duluth Complex: Implications for metamorphic and metasomatic processes associated with mafic magma chambers. *Geochim. Cosmochim. Acta* **65**, 2965–2978 (2001).
- Schneider, D. A., Bickford, M. E., Cannon, W. F., Schulz, K. J. & Hamilton, M. A. Age of volcanic rocks and synpositional iron formations, Marquette Range Supergroup: implications for the tectonic setting of Paleoproterozoic iron formations of the Lake Superior region. *Can. J. Earth Sci.* **39**, 999–1012 (2002).
- Masaitis, V. L. *et al.* in *Large Impacts and Planetary Evolution* (eds Dressler, B. O. & Sharpton, V. L.) 317–321 (Geological Society of America, Special Paper Vol. 339, 1999).
- Grieve, R. A. F. & Cintala, M. J. An analysis of differential impact melt-crater scaling and implications for the terrestrial impact record. *Meteoritics* **27**, 526–538 (1992).
- Pierazzo, E., Vickery, A. M. & Melosh, H. J. A reevaluation of impact melt production. *Icarus* **127**, 408–423 (1997).
- Ivanov, B. A. & Deutsch, A. in *Large Impacts and Planetary Evolution* (eds Dressler, B. O. & Sharpton, V. L.) 389–396 (Geological Society of America, Special Paper Vol. 339, 1999).
- Therriault, A. M., Fowler, A. D. & Grieve, R. A. F. The Sudbury Igneous Complex: a differentiated impact melt-sheet. *Econ. Geol.* **97**, 1521–1540 (2002).
- Ames, D. E., Golightly, J. P., Lightfoot, P. C. & Gibson, H. L. Vitric compositions in the Onaping Formation and their relationship to the Sudbury Igneous Complex, Sudbury Structure. *Econ. Geol.* **97**, 1541–1562 (2002).
- Ames, D. E., Watkinson, D. H. & Parrish, R. R. Dating of a regional hydrothermal system induced by the 180 Ma Sudbury impact event. *Geology* **26**, 447–450 (1998).
- Coleman, A. P. The Sudbury nickel deposits. *Ontario Bureau Mines Rep.* **1902**, 253–303 (1903).
- Lightfoot, P. C., Keays, R. R., Morrison, G. G., Bite, A. & Farrell, K. P. Geochemical relationships in the Sudbury Igneous Complex: origin of the main mass and Offset dikes. *Econ. Geol.* **92**, 289–307 (1997).
- Cowan, E. J. & Schwerdtner, W. M. in *Proceedings of the Sudbury-Noril'sk Symposium* (eds Lightfoot, P. C. & Naldrett, A. J.) 45–56 (Ontario Geological Survey, Special Vol. 5, 1994).
- Richardson, T. & Burnham, O. M. Precious metal analysis at the Geoscience Laboratories: results from the new low-level analytical facility. *Ont. Geol. Surv. Open File Rep.* **6100**, 35 (2002).
- Dickin, A. P., Nguyen, T. & Crockett, J. H. in *Large Impacts and Planetary Evolution* (eds Dressler, B. O. & Sharpton, V. L.) 361–371 (Geological Society of America, Special Paper Vol. 339, 1999).
- Mungall, J. E. *Program with Abstracts for 38th Annual GSA Northeastern Division Meeting, Halifax* 23–29 (Geological Society of America, 2003).
- Lightfoot, P. C. & Farrow, C. E. G. Geology, geochemistry, and mineralogy of the Worthington Offset dike: a genetic model for Offset dike mineralization in the Sudbury Igneous Complex. *Econ. Geol.* **97**, 1419–1446 (2002).
- Wedepohl, K. H. The composition of the continental crust. *Geochim. Cosmochim. Acta* **59**, 1217–1232 (1995).
- Gao, S. *et al.* Chemical composition of the continental crust as revealed by studies in East China. *Geochim. Cosmochim. Acta* **62**, 1959–1975 (1998).
- Pope, K. O., Kieffer, S. W. & Ames, D. E. Empirical and theoretical comparisons of the Chicxulub and Sudbury Impact Craters. *Meteorit. Planet. Sci.* **39**, 97–116 (2004).
- Campins, H. & Swindle, T. D. Expected characteristics of cometary meteorites. *Meteorit. Planet. Sci.* **33**, 1201–1211 (1998).
- McDonough, W. F. & Sun, S.-s. The composition of the Earth. *Chem. Geol.* **120**, 223–253 (1995).
- Pařava, J., Barnes, S.-J. & Vymazalová, A. The use of mantle normalization and metal ratios in the identification of the sources of platinum-group elements in various metal-rich black shales. *Mineral. Dep.* **38**, 775–783 (2003).
- Alvarez, L. W., Alvarez, W., Asaro, F. & Michel, H. V. Extraterrestrial cause for the Cretaceous–Tertiary boundary extinction. *Science* **208**, 1095–1108 (1980).
- Colodner, D. C., Boyle, E. A., Edmond, J. M. & Thomson, J. Post-depositional mobility of platinum, iridium and rhenium in marine sediments. *Nature* **358**, 402–404 (1992).
- Heymann, D. *et al.* in *Large Impacts and Planetary Evolution* (eds Dressler, B. O. & Sharpton, V. L.) 345–360 (Geological Society of America, Special Paper Vol. 339, 1999).

Acknowledgements This study was generously supported by the University of Toronto, by the Geoscience Laboratories of the Ontario Geological Survey, and by the Geological Survey of Canada.

Author's contributions Mapping and sampling was performed by J.E.M., D.E.A. and J.J.H.; geochemical interpretation was done by J.E.M., D.E.A. and J.J.H., in decreasing order of relative contribution.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to J.E.M. (mungall@geology.utoronto.ca).

Doubling the estimate of invertebrate biomass in a rainforest canopy

Martin D. F. Ellwood & William A. Foster

University Museum of Zoology, Department of Zoology, University of Cambridge, Downing Street, Cambridge CB2 3EJ, UK

Forest canopies represent the functional interface between 90% of the Earth's terrestrial biomass and the atmosphere¹ and include some of the most threatened of all terrestrial ecosystems². However, we lack even a basic understanding of how the biomass of plants and animals is distributed throughout forest canopies, even though this information is vital for estimating energy flow, carbon cycling, resource use and the transfer of materials within this ecosystem^{3,4}. Here we measure the biomass of invertebrates living in a common rainforest epiphyte, describe a striking relationship between fern size and the biomass of animals within the ferns, and reveal that one large epiphyte may contain an invertebrate biomass similar to that found in the whole of the rest of the tree crown on which it is growing. Using these data, we show that including the fauna of these epiphytes—a neglected component in rainforest ecosystems—can more than double our estimate of the total invertebrate biomass in an entire rainforest canopy.

Epiphytes are known to be important to the structure and function of rainforest canopies, but there have been no detailed studies of how the animals living within epiphytes might contribute to the canopy ecosystem. The invertebrate biomass contained within bromeliads has been estimated for selected forests on an island in the Caribbean⁵, but no previous work has estimated the contribution of an epiphyte fauna to the invertebrate biomass of the canopy as a whole. Here, we chose the bird's nest fern, *Asplenium nidus* (Fig. 1a) as our model epiphyte, because it is abundant^{6–8}, it occurs at all levels within the canopy, it can grow to exceptionally large sizes (~200 kg fresh weight) and it is widespread within the palaeotropics⁹.

We collected the entire fauna (arthropods, gastropods, annelids, amphibians, reptiles and mammals) from 35 ferns of a range of sizes (1–200 kg fresh weight) and vertical heights (2–52 m) from 7 ha of lowland dipterocarp rainforest in Borneo (see Methods). We took the ferns down and sampled the whole of each fern—not just the litter trapped within it. There was a striking relationship between fern size and the biomass of invertebrates within the ferns (Fig. 1b); smaller ferns contained just a few grams of animal biomass, but this increased exponentially in ferns with more than 30 leaves, reaching 158 g in the largest fern.

Because the larger ferns contained disproportionately more animal biomass, these ferns probably make a greater contribution to the canopy biomass as a whole. We therefore looked at large ferns in more detail and selected five specimens, each of which was growing in a separate crown of the dipterocarp *Parashorea tomentella* that contained no other large epiphytes. We sampled the tree crowns by fogging them with insecticide¹⁰. Because fogging does not collect vertebrate animals, we limited our attention to the invertebrates. A total of 205,428 invertebrate animals were collected from the five ferns, representing an average total biomass of 88 ± 14 g (dry weight) per fern. The invertebrate biomass within the ferns was made up of a diverse range of taxa (Fig. 2a). The ferns contained a total of 27 orders compared with a total of 26 orders in the crowns. However, the bulk of the invertebrate biomass in the ferns was made up of fewer taxa, compared with the more even biomass distribution among the taxa in the crowns (Fig. 2b).

We directly compared the invertebrate biomass in the ferns and

crowns using a precision-fogging technique that sampled each fern and an equivalent area of crown just above it¹¹. The invertebrate biomass was two orders of magnitude higher in the fern samples than in the crown samples ($42 \pm 6 \text{ g m}^{-2}$ versus $0.35 \pm 0.04 \text{ g m}^{-2}$; paired *t*-test: $t = 7.31$, $P = 0.005$, $n = 5$). The higher biomass of invertebrates in the ferns was not only due to larger numbers of individuals, but to a combination of increased abundance and increased body size: the invertebrates in the ferns were both more abundant and larger than those in the crowns. Of the ten largest-bodied orders, eight contained individuals that were significantly heavier in the ferns than those in the crowns (Fig. 3). This direct comparison of the ferns with the crowns also enabled us to estimate the effectiveness of fogging as a method of sampling the fern fauna; fogging sampled $53 \pm 8\%$ of the invertebrate biomass within the ferns.

To assess the relative contribution of the fern fauna biomass, we compared the invertebrate biomass inside the large ferns with that inside the entire crowns of the trees on which the ferns were growing. Before removing the ferns, we fogged 21 m^2 of the crown of each tree and measured the horizontal cross-sectional area of each crown. The product of the invertebrate biomass per square metre in each fog ($\text{mean} = 0.21 \pm 0.04 \text{ g m}^{-2}$, $n = 5$) and the estimated surface area of each crown ($\text{mean} = 361 \pm 28 \text{ m}^2$, $n = 5$) gave us a measure of the total invertebrate biomass—

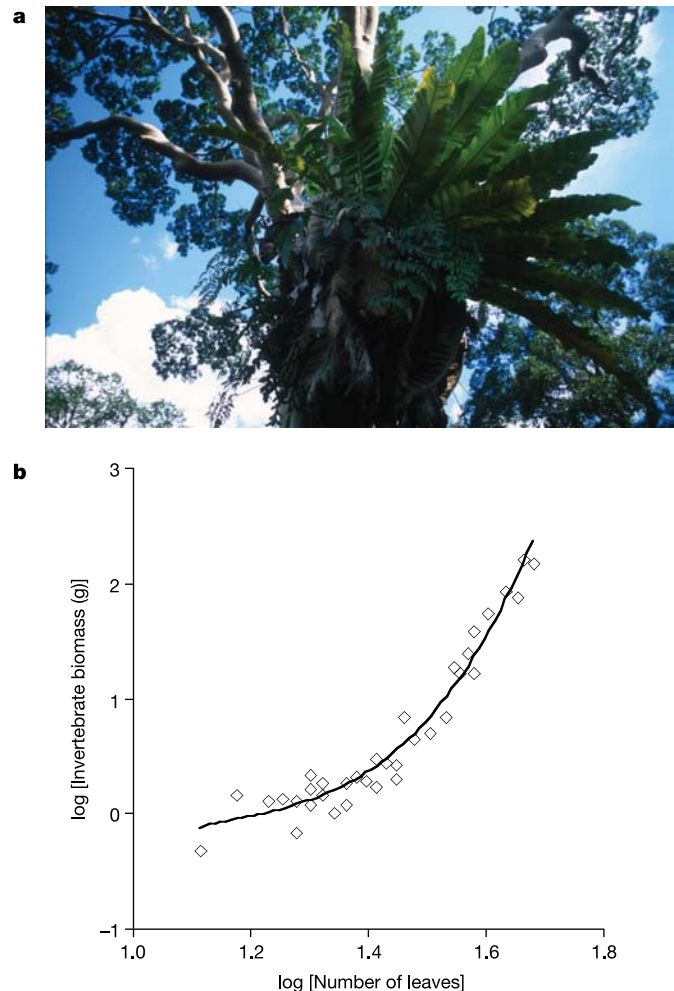


Figure 1 Invertebrate biomass in bird's nest ferns. **a**, The bird's nest fern (*A. nidus*) with leaves reaching up to 2 m in length. **b**, Relationship between the total invertebrate biomass found inside 35 ferns and the number of leaves on each of the ferns ($y = 13.9x^3 - 48.8x^2 + 58.1x - 23.5$, $r^2 = 0.96$).

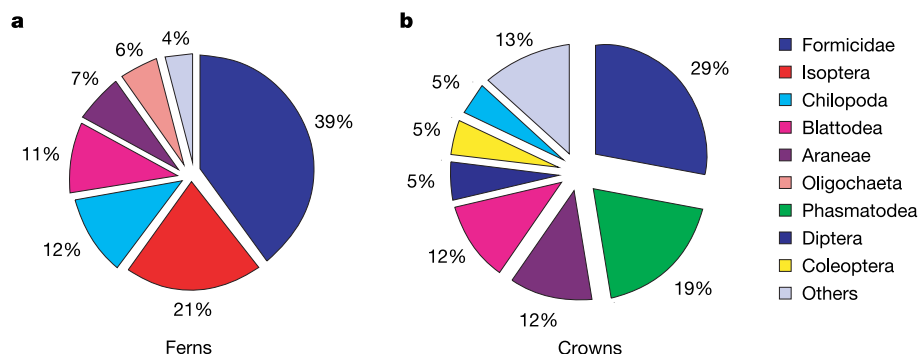


Figure 2 Contributions of various taxa to overall invertebrate biomass in ferns and crowns. **a**, Proportional contributions of the taxa that contributed most to the invertebrate biomass in five large *Asplenium* ferns (22–26 kg dry weight) (total invertebrate biomass = 439 g).

b, Proportional contributions of the taxa that contributed most to the invertebrate biomass in five tree crowns of *P. tomentella* (total invertebrate biomass = 21 g).

excluding the fern fauna—in the crown of each tree. This figure (86 ± 18 g) is of the same order of magnitude as the total mean invertebrate biomass in a single large fern (88 ± 14 g). Fogging will, of course, sample only a proportion of the invertebrates in the tree crowns. If this proportion is the same as that for the ferns (53%), then the true total in the crowns would be about 160 g, but this is likely to be an overestimate, because the crowns—unlike the ferns—are not designed to capture and retain plant and animal matter. Our results therefore suggest that there is almost as much invertebrate biomass in a single fern as in the whole of the rest of the tree crown on which it is growing.

To quantify the overall contribution of these epiphytes to the invertebrate biomass of the entire rainforest canopy, we obtained independent estimates on a hectare basis of the biomass contained within the ferns and within the rest of the crowns. To measure the fern component, we needed to estimate the number and size of the ferns in a hectare and the invertebrate biomass within each fern. We used our collection of 35 ferns of different sizes to produce a regression equation that described the total invertebrate biomass in a fern on the basis of its number of leaves (Fig. 1b). Using this equation and the results from intensive fern surveys in 7 ha of forest, we estimated that the total amount of invertebrate biomass contained within the ferns is $3776 \pm 275 \text{ g ha}^{-1}$ ($n = 7$). To estimate the invertebrate biomass per hectare of tree crowns, we fogged the canopies of five trees of *P. tomentella* using 21 trays of 1 m^2 each per tree. We extrapolated these measurements to a hectare basis to provide an estimate of tree crown invertebrate biomass of $2,026 \pm 373 \text{ g ha}^{-1}$ ($n = 5$). If we make the rather conservative assumption that fogging extracts the same proportion of invertebrate biomass from tree crowns as it does from ferns—that is, 53%—then the invertebrate biomass within the tree crowns could be as high as $3,822 \pm 703 \text{ g ha}^{-1}$.

Although we fogged only one species of tree, the numbers and biomass of animals that we collected are in agreement with data from other studies of canopy arthropods in southeast-Asian lowland-tropical rainforest and Amazonia^{12–15}. One of the most comprehensive fogging studies of canopy invertebrates was carried out in the neighbouring island of Sulawesi¹², which sampled a range of tree species from a large area of lowland forest, using a total of 960 trays (each 1 m^2). This study gave a mean abundance of 213 animals (range 64–461) per square metre, compared with our mean abundance measure of 160 (range 78–553). The Sulawesi study did not give data on biomass, but by using our data on the mean lengths of individuals of canopy invertebrates, we can estimate that the mean biomass in the Sulawesi study was 210 (range 46–563) mg m^{-2} compared with the figure from the current study of 214 (range 34–618).

Our results show that this one species of epiphyte can contain half

the entire invertebrate biomass within a hectare of rainforest canopy. This indicates that a substantial proportion of the energy and nutrients converted into animal biomass in this rainforest canopy is being funnelled through this one species of fern. We have almost certainly underestimated the general role of large epiphytes in the canopy ecosystem. *A. nidus* is just one species of large epiphyte; in this part of Borneo alone there are another 12 species of *Asplenium*⁷ and some of these are equal in size to *A. nidus*. Other large ferns such as the stag's horn fern (*Platycerium* spp.) or flowering epiphytes of the Pandanaceae could well contain similar amounts of invertebrate biomass. It is worth noting that trash-basket ferns of the genus *Asplenium* are widespread and abundant in rainforests throughout the palaeotropics. Because of their size and abundance, it is likely that these particular ferns have a marked effect on invertebrate biomass distribution in rainforest canopies over a large geographical area.

To our knowledge, this is the first estimate of the contribution of epiphytes to the total invertebrate biomass of an entire rainforest canopy. Our results indicate that epiphytes have a more substantial impact on the animal communities in tropical rainforests than was previously imagined and challenge our current understanding of

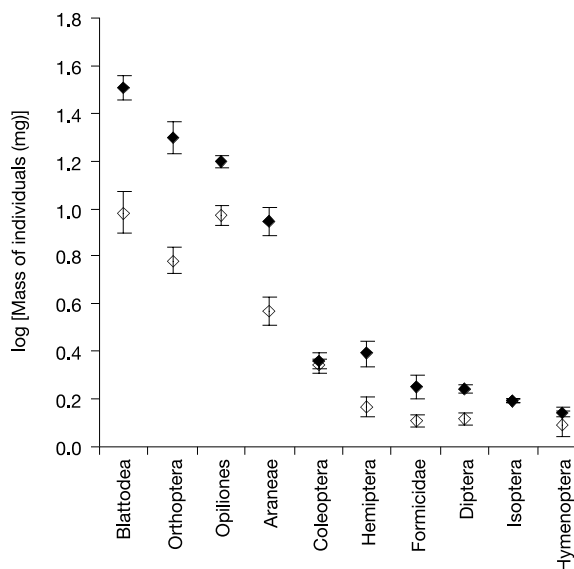


Figure 3 Individual biomass. The mean mass of individual animals from eight out of ten of the largest-bodied taxa was greater in the five large ferns (closed symbols) than in the corresponding host crowns (open symbols) ($F_{9,80} = 59.3$, $P < 0.001$). There was no significant interaction between individual mass and habitat ($F_{9,80} = 1.49$, $P = 0.172$).

the distribution and functional significance of animals in the world's rainforest canopies. Two further kinds of investigation are required to establish the broader significance of these findings. First, is the occurrence of substantial animal biomass in the epiphytes of rainforest canopies a widespread phenomenon? Do other large epiphytes of the high canopy, both in southeast Asia and in other tropical regions, contribute significantly to the total animal biomass of the canopy? The only way to answer this involves the challenging and laborious process of bringing these epiphytes down from the canopy and sampling them completely. Second, what role does this animal biomass play in the functioning of the canopy ecosystem as a whole? What function do the animals have in carbon and nutrient cycling, in decomposition and in the maintenance of biodiversity in the canopy? This can only be solved by detailed sampling and by conducting experiments within the canopy itself.

There will undoubtedly be other, overlooked concentrations of invertebrates in the canopy, living in a wide range of habitats other than epiphytes^{16,17}. As further studies are made of the animals both in epiphytes and elsewhere in the canopy, the estimates of total invertebrate biomass will increase and the relative importance of epiphytes may need to be revised. This paper is the first step in this process and shows that the inclusion of one species of epiphyte can have dramatic effects on the estimates of invertebrate biomass in the canopy. □

Methods

Fogging

We used Pybuthrin 33, a synthetic pyrethroid insecticide with low mammalian toxicity, in a Swing Fog SN 50 fogging machine. Circular trays (1 m²) made of waterproof material and aluminium hoops were used to collect falling invertebrates. Fogging began at daybreak and lasted approximately 12 minutes. To determine the horizontal cross-sectional area of each crown we measured the distance at 20° intervals from the base of the trunk to the tips of the branches.

Fern collecting

Ferns were collected from undisturbed lowland dipterocarp rainforest in Danum Valley, Sabah, Malaysian Borneo (4° 58' N, 117° 48' E). Fern collection took place in approximately 100 ha of forest adjacent to the Danum Valley field centre and the Segama river. Each fern was netted before removal and carried back to the laboratory, where it was carefully dissected and the animals extracted using conventional methods^{11,18,19}. Plant material was hand-sorted after the animals had been extracted. Animal material was then examined using binocular dissecting microscopes capable of distinguishing animals larger than 50 µm. We were thus able to identify all arthropods, including small animals such as mites and springtails.

Fern surveys

The seven 1-ha plots used for the fern surveys were located in the same 100 ha of undisturbed forest from where we collected the ferns. Using binoculars, we performed 10-m transects at ground level and treetop surveys from an emergent tree in the centre of each plot.

Biomass estimation

Animal biomass was calculated from length measurements of individual animals^{5,20,21}. We established the relationship between leaf number and animal biomass (Fig. 1b) using 35 ferns collected from the same 100 ha of undisturbed forest mentioned above. From this relationship, after counting the number of ferns in each hectare and the number of leaves on each of the ferns, we were able to estimate the total animal biomass in the ferns of 1 ha of rainforest. The estimate for the crowns was based on the mean animal biomass in 21 trays of 1 m² each in the crown fogs from each of five trees, multiplied up to a hectare basis.

Received 12 March; accepted 14 April 2004; doi:10.1038/nature02560.

- Ozanne, C. M. P. *et al.* Biodiversity meets the atmosphere: a global view of forest canopies. *Science* **301**, 183–186 (2003).
- Stork, N. E. Insect diversity: facts, fiction and speculation. *Biol. J. Linn. Soc.* **35**, 321–337 (1988).
- Brown, J. H. & Maurer, B. A. Body size, ecological dominance and Cope's rule. *Nature* **324**, 248–250 (1986).
- Pagel, M. D., Harvey, P. H. & Godfray, H. C. J. Species-abundance, biomass, and resource-use distributions. *Am. Nat.* **138**, 836–850 (1991).
- Richardson, B. A., Richardson, M. J., Scatena, F. N. & McDowell, W. H. Effects of nutrient availability and other elevational changes on bromeliad populations and their invertebrate communities in a humid tropical forest in Puerto Rico. *J. Trop. Ecol.* **16**, 167–188 (2000).
- Ellwood, M. D. E., Jones, D. T. & Foster, W. A. Canopy ferns in lowland dipterocarp forest support a prolific abundance of ants, termites and other invertebrates. *Biotropica* **34**, 575–583 (2002).
- Parris, B. S. A provisional check-list of pteridophytes from the Danum Valley, Sabah. *Sandakania* **9**, 77–85 (1997).

- Pocs, T. The epiphytic biomass and its effect on the water balance of two rain forest types in the Uluguru Mountains (Tanzania, East Africa). *Acta Bot. Acad. Sci. Hung.* **26**, 143–167 (1980).
- Benzing, D. H. *Vascular Epiphytes* 354 (Cambridge Univ. Press, Cambridge, 1990).
- Adis, J., Basset, Y., Floren, A., Hammond, P. M. & Linsenmair, K. E. Canopy fogging of an overstory tree—recommendations for standardization. *Ecotropica* **4**, 93–97 (1998).
- Ellwood, M. D. E. & Foster, W. A. in *The Global Canopy Handbook: Techniques of Access and Study in the Forest Roof* (eds Mitchell, A. W., Secoy, K. R. J. & Jackson, T.) 115–119 (Glob. Canopy Prog., Oxford, 2002).
- Stork, N. E. & Brendell, M. J. D. in *Insects and the Rain Forests of South East Asia* (Wallacea) (eds Knight, W. J. & Holloway, J. D.) 173–190 (R. Entomol. Soc., London, 1990).
- Stork, N. E. The composition of the arthropod fauna of Bornean lowland rain forest trees. *J. Trop. Ecol.* **7**, 161–180 (1991).
- Stork, N. E. & Brendell, M. J. D. in *Natural History of Seram, Maluku, Indonesia* (eds Edwards, I. D., Macdonald, A. A. & Proctor, J.) 115–130 (Intercept, Andover, 1993).
- Guerrero, J. C. H., da Fonseca, C. R. V., Hammond, P. M. & Stork, N. E. in *Arthropods of Tropical Forests: Spatio-Temporal Dynamics and Resource Use in the Canopy* (eds Basset, Y., Novotny, V., Miller, S. E. & Kitching, R. L.) 170–175 (Cambridge Univ. Press, Cambridge, 2003).
- Basset, Y. Invertebrates in the canopy of tropical rain forests: how much do we really know? *Plant Ecol.* **153**, 87–107 (2001).
- Nadkarni, N. M. & Longino, J. T. Invertebrates in canopy and ground organic matter in a Neotropical montane forest, Costa Rica. *Biotropica* **22**, 286–289 (1990).
- Besuchet, C., Burckhardt, D. & Lobl, I. The Winkler/Moczarski elector as an efficient extractor for fungus and litter Coleoptera. *Coleopterists' Bull.* **41**, 392–394 (1987).
- Ellwood, M. D. E. & Foster, W. A. Line insertion techniques for the study of high forest canopies. *Selbyana* **22**, 97–102 (2001).
- Rogers, L. E., Hinds, W. T. & Buschbom, R. L. A general weight vs. length relationship for insects. *Ann. Entomol. Soc. Am.* **69**, 387–389 (1976).
- Schoener, T. W. Length-weight regressions in tropical and temperate forest-understorey insects. *Ann. Entomol. Soc. Am.* **73**, 106–109 (1980).

Acknowledgements We thank Clare College, the Department of Zoology and the Museum of Zoology of the University of Cambridge, Yayasan Sabah, Danum Valley Management Committee and The Economic Planning Unit of the Prime Minister's Department, Kuala Lumpur. We thank our collaborator in Sabah, Chey V. K. For comments on the manuscript, we thank N. Davies, N. Mundy and M. Brooke. The photograph of *A. nidus* (Fig. 1a) was provided by courtesy of R. Dial. We are grateful to M. and D. Ellwood for their assistance in the field, and to those who helped sort insects. M.D.E.E. was funded by the Natural Environment Research Council (NERC). This research was carried out while M.D.E.E. was a participant in the Royal Society's south-east Asian rain forest research programme (Publ. No. A405).

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to M.D.E.E. (mdfe2@cam.ac.uk).

Unusually dynamic sex roles in a fish

Elisabet Forsgren^{1,2}, Trond Amundsen², Åsa A. Borg^{2*} & Jens Bjelvenmark¹

¹Department of Marine Ecology, Göteborg University, Kristineberg Marine Research Station, SE-45034 Fiskebäckskil, Sweden

²Department of Biology, Norwegian University of Science and Technology, NO-7491 Trondheim, Norway

* Present address: Department of Biology, University of Oulu, Box 3000, FIN-90014 Oulu, Finland

Sex roles are typically thought of as being fixed for a given species. In most animals males compete for females, whereas the females are more reluctant to mate. Therefore sexual selection usually acts most strongly on males^{1,2}. This is explained by males having a higher potential reproductive rate than females, leading to more males being sexually active (a male-biased operational sex ratio)^{3,4}. However, what determines sex roles and the strength of sexual selection is a controversial and much debated question^{3,5–10}. In this large-scale field study, we show a striking temporal plasticity in the mating competition of a fish (two-spotted goby, *Gobiulus flavescens*). Over the short breeding season fierce male–male competition and intensive courtship behaviour in males were replaced by female–female competition and actively courting females. Hence, sex role reversal occurred rapidly. This is the first time that a shift in sex roles has been

shown in a vertebrate. The shift might be explained by a large decline in male abundance, strongly skewing the sex ratio towards females. Notably, the sex role reversal did not occur at an equal operational sex ratio, contrary to established sex role theory^{3,4}.

Animal sex roles are termed 'conventional' when male competition for mates is strongest and 'reversed' when female competition is strongest¹¹. Parental care and choosiness in mate choice are not included in this definition of sex roles. Hence, the term 'sex roles' as used here has a more specific meaning than it might in everyday language. Mating competition can be expressed through agonistic behaviour that prevents conspecifics from mating, or through courtship to attract the opposite sex. The intensity of mating competition is largely explained by mate availability, expressed by the operational sex ratio (OSR); that is, the ratio of ready-to-mate males to ready-to-mate females^{3,4}. If the OSR varies within a species, spatially or temporally, the intensity of mating competition should vary accordingly, as has been shown in several taxa⁴. If the variation in the OSR is large enough, a change in sex roles is expected. The only known case of variable sex roles is found in katydid insects, in which food abundance determines which sex is most sexually active^{12–14}. It is not known how widespread flexible sex roles are, or whether vertebrates also show such plasticity. No previous study has tested the central prediction of traditional sex role theory; that is, that sex roles should shift at an equal OSR. This may be partly due to difficulties in measuring the OSR in natural populations^{3,6}, and partly because there are very few documented cases where variation in the OSR is sufficient to allow both conventional and reversed sex roles within a species. Such variation is little studied, however, and might not necessarily be rare.

In the present study of a wild population of the two-spotted goby, we estimated the OSR (see Methods and Table 1) over the breeding season at several localities, and tested whether the strength and direction of mating competition followed the variation in the OSR as predicted^{3,4}. We observed a large (tenfold) decline in the number of males over the season, whereas female numbers did not vary significantly (Fig. 1). Consequently, the adult sex ratio changed from unbiased to female biased (repeated measures analysis of variance (ANOVA) on arcsine($\sqrt{x}$)-transformed data; $F_{3,9} = 11.37$, $P < 0.0001$). The decline in male abundance is likely to be due to a higher mortality among males than females during the breeding season. A high male mortality might be caused by exhaustion resulting from parental care¹⁵ or male–male competition¹⁶, a higher susceptibility to parasites and diseases¹⁷, or an increased predation risk¹⁸.

The decline in male abundance (Fig. 1 and Table 1) was accompanied by a marked increase in the proportion of round

(that is, ready to mate) females over the season (repeated measures ANOVA on arcsine($\sqrt{x}$)-transformed data; $F_{3,9} = 5.02$, $P = 0.007$, Table 1). Consequently, there was a change in the OSR from heavily male biased early in the season to heavily female biased by the end (Fig. 2). Late in the season mating options for the many mature females were limited. The observed change in the OSR was contrary to predictions made based on temperature-dependent effects on male and female potential reproductive rates in fishes with paternal care^{19,20}. Hence, our study emphasizes the importance of factors affecting the OSR not through reproductive rates, but directly through the abundance of the two sexes. Declines in the number of males, leading to female-biased sex ratios late in the breeding season, have been reported in other fish²¹. In such cases sex role reversals should be expected. Similarly, sex ratios may vary spatially²², or as a consequence of sex-biased fisheries²³. Thus, conditions favouring flexible sex roles may not be uncommon.

As theory predicts^{3,4}, the change in the OSR was accompanied by a marked change in mating competition over the season, reflected in agonistic behaviour and courtship displays. The proportion of focal males (that is, nest-holding males observed; see Methods) that were involved in agonistic interactions with other males declined sharply over the season ($\chi^2 = 44.05$, $n = 177$, $P < 0.0001$), as did the number of agonistic acts performed by each focal male (Kruskal–Wallis ANOVA; $H = 38.73$, $n = 177$, $P < 0.0001$). This was not only owing to fewer male–male encounters as the season progressed, but also because focal males showed a much reduced propensity to behave agonistically at each encounter with another male (Fig. 3a). Agonistic acts between females showed the opposite temporal pattern to that of males. Over the season, female agonistic behaviour occurred near to an increasing proportion of focal males ($\chi^2 = 43.61$, $n = 176$, $P < 0.0001$), indicating female competition for access to males. The number of female agonistic interactions observed near to focal males also increased over the season (Kruskal–Wallis ANOVA, $H = 44.84$, $n = 176$, $P < 0.0001$). In contrast to males, females showed an increasing propensity to behave agonistically as the season progressed (Fig. 3a). Thus, over the season, male–male competition was replaced by female–female competition.

An equally marked change was found for courtship behaviour. Early in the season, males actively courted females, but later on this courting almost entirely ceased. This was shown by a decline in both the proportion of males that courted ($\chi^2 = 12.61$, $n = 177$, $P = 0.006$) and the number of courtship displays performed by

Table 1 Variables determining OSR variation of two-spotted gobies over the season

Variable	Sampling period*			
	1	2	3	4
Territorial (stationary) males				
Number per m ²	0.56	0.32	0.13	0.07
Percentage of all males	61.7	60.2	53.3	73.2
Spawning-ready (round) females†				
Number per m ²	0.15	0.39	0.29	0.33
Percentage of all females	37.3	59.3	78.1	75.4
Egg cover in artificial nest sites in the wild				
Percentage of nest area	29.7	40.3	64.7	76.5
Number of additional clutches possible per nest‡	2.98	2.53	1.49	0.99

All values are means for the respective sampling periods. See Methods for details of calculating the OSR.

*Each locality was visited four times over the breeding season (May–July). See Methods for exact dates.

†Includes females with roundness scores of 2 and 3 (ready to mate).

‡The 'male availability factor'; see Methods.

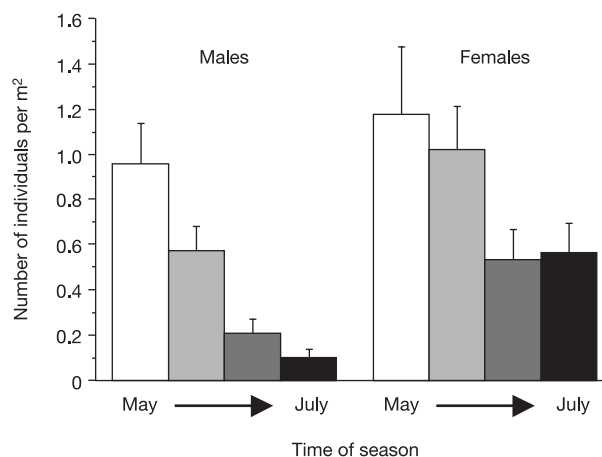


Figure 1 Abundance of two-spotted gobies over the breeding season. Mean ($\pm$ s.e.m.) number of males per m² (left; repeated measures ANOVA; $F_{3,9} = 20.46$, $P < 0.0001$) and females per m² (right; $F_{3,9} = 2.69$, $P = 0.07$) observed along transects ($n = 10$ localities) in relation to season (four sampling occasions from May to July).

each male (Kruskal–Wallis ANOVA; $H = 13.39$, $n = 176$, $P = 0.004$). The decreasing propensity of focal males to court visiting females, as the season progressed, was even more pronounced (Fig. 3b). In contrast, females hardly ever courted early in the season but became very active in courtship later on, as shown by an increase in the proportion of focal males courted ($\chi^2 = 42.52$, $n = 177$, $P < 0.0001$). Each male received more female courtship acts later in the season (Kruskal–Wallis ANOVA, $H = 49.74$, $n = 177$, $P < 0.0001$) and the propensity of visiting females to court the focal male also increased sharply over the season (Fig. 3b). Late in the season, individual males were often surrounded by up to 20 round females courting them at close range. Thus, there was a reversal of courtship roles over the season. Reversed courtship roles associated with nest site shortage^{24,25} or time of season²⁶ have also been documented in other fishes. However, as courtship is just one facet of mating competition, it is unknown whether sex role reversals occurred in these cases; that is, whether females and not males faced the strongest mating competition. In order to show conclusively a shift in sex roles, agonistic and courtship behaviour should change together, as found here.

Our results show a shift in sex roles, from conventional to reversed, in only a couple of months. Early in the season, males competed aggressively with each other for matings and were very active in courtship, whereas late in the season females behaved agonistically and took over as the courting sex. Such a shift in both agonistic and courtship behaviour of both sexes has, to our knowledge, never been shown in any animal species. In two-spotted gobies the strength of sexual selection should be greater on males at the beginning of the season but greater on females later on. Whether there is a corresponding change in which sex is the most selective in mate choice²⁷ remains to be tested; however, we know that both sexes can exhibit mating preferences in this species (ref. 28; Å.A.B., T.A. and E.F., unpublished data). Plasticity of which sex is most choosy has been found in katydid insects^{12,13}. Choosiness is an important facet of sexual behaviour, the dynamics of which deserve further study.

The direction and strength of mating competition roughly followed the change in the OSR over time, as predicted⁴. This supports the notion that the OSR is an important determinant of mating competition and sex roles. However, a closer examination of the OSR at the point of sex role reversal does not support classical sex role theory, which predicts that sex roles shift at an equal OSR⁴.

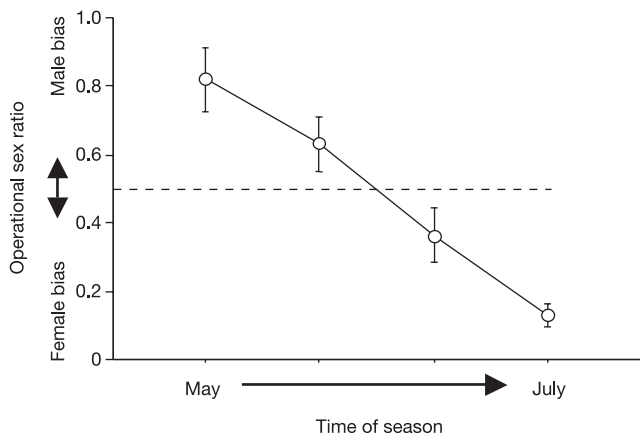


Figure 2 Seasonal change in the OSR. Estimated OSR (mean $\pm$ s.e.m.) of two-spotted gobies over the breeding season (repeated measures ANOVA on arcsine($\sqrt{x}$)-transformed data; $F_{3,9} = 18.13$, $P < 0.0001$). Only stationary males and round females observed along transects ($n = 10$) are included. The fact that each male can care for the eggs of more than one female is taken into account (see Methods). The dashed line at 0.5 indicates when sex roles should shift according to established sex role theory.

The pattern of agonistic and courtship behaviour showed that sex role reversal occurred between the third and fourth sampling period (Fig. 3); that is, at an OSR between 0.36 and 0.13 (Fig. 2). Hence, when the sex roles shifted the OSR was not equal (0.5) but was female biased. Recent, and so far untested, life-history-based theoretical models^{9,27} suggest that the sex-specific costs of breeding are fundamental in determining sex roles and the strength of sexual selection, as the sex paying the highest reproductive costs should be less likely to compete over matings. Sex role shifts may therefore occur at any OSR but will only rarely occur at 0.5 (refs 9, 27). According to the life-history-based models, a sex role shift in a female-biased situation predicts a higher reproductive cost for females than for males. Whether this is true for two-spotted gobies remains to be tested. Providing empirical data on when sex role shifts occur, our study supports the assertion that the OSR is not the sole predictor of sex roles^{9,27}. Notably, it is unknown how quickly animals can respond to changes in the OSR. It is conceivable that information gathering and/or hormonal processes produce a lag in the animals' response, leading to a delay in the sex role shift relative to the theoretical optimum. Even if field studies like ours do inevitably estimate the OSR with some error, such difficulties should not prevent researchers from studying how sex roles relate to the OSR in the wild^{3,4}.

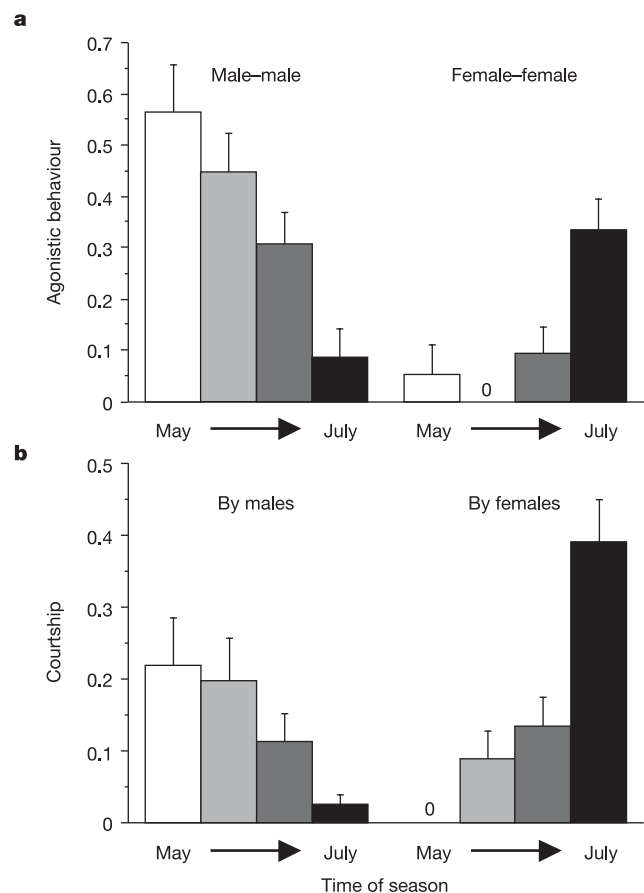


Figure 3 Change in male and female mating competition over the breeding season in the two-spotted goby. **a**, Propensity to behave agonistically when encountering same-sex individuals near territorial (focal) males. Left, the proportion of visiting males that the focal male behaved agonistically towards (Kruskal–Wallis ANOVA; $H = 24.27$, $P < 0.0001$, $n = 125$). Right, the proportion of visiting females that behaved agonistically towards other simultaneously visiting females ($H = 25.77$, $P < 0.0001$, $n = 89$). **b**, Propensity to perform courtship. Left, the proportion of visiting females that the focal male courted ($H = 13.71$, $P = 0.003$, $n = 145$). Right, the proportion of visiting females that courted the focal male ($H = 43.61$, $P < 0.0001$, $n = 145$). Bars represent mean $\pm$ s.e.m.

Our study reveals a flexibility in vertebrate sex roles, which apparently can shift quickly within the same population. One implication is that it can sometimes be difficult, or even misleading, to classify species as either conventional or role-reversed. Variation in the factors that determine sex roles has been little studied; dynamic sex roles may thus be more widespread than currently known. This would have important implications for sexual selection processes. In particular, females might be subject to sexual selection more often than previously realized. To fully understand sexual selection in a species, it is essential to explore the range of variation in selection regimes, be it spatial or temporal. □

Methods

Study species

The two-spotted goby is a small, shoaling fish that is very common along the rocky shores of western Europe. It typically lives for one year only. During spring and summer it inhabits the shallow algal zone, breeding from May to July. Non-breeding individuals often form semipelagic foraging shoals, which are sometimes very large. Breeding males defend nest sites in brown algae or empty mussels, to which they attract females for spawning. During spawning, eggs are laid in a single layer in the nest. Males provide uniparental care (which involves fanning, cleaning and defence) until the eggs hatch. A single male can care for clutches from several females simultaneously. Males and females can reproduce repeatedly over the breeding season. During breeding, males have iridescent blue markings on their body and fins, whereas females develop bright yellow/orange bellies²⁸.

General procedures

Data were collected at ten different localities in the Gullmarsfjord, western Sweden (58° 15' N, 11° 28' E), each of which was visited on four different occasions during the breeding season of 1999 (24–28 May, 7–9 June, 28 June–1 July and 19–21 July). At each location we marked a transect (18–33 m long) along the rocky shoreline within typical breeding habitat, which was dominated by algal vegetation. While slowly snorkelling along the transects, we recorded the number of males and females encountered within a 1-m-wide strip along the length of the transect. Fish were recorded to a depth of 1 m (less if shallower), and were typically solitary or in small groups. When foraging shoals (see above) were encountered in the transects, they were included in the estimation of male and female abundance, but not in the OSR calculation (see below). About 20% of recorded fish were found in such shoals. When these shoals did not allow sex identification of all members, we estimated male and female numbers from the recorded sex ratio of sexed shoaling fish at the same sampling period. Two observers swam along each transect on each occasion; mean values were used in analyses.

Operational sex ratio

Along the transects we scored the roundness of females (1, slim/slightly round; 2, clearly round; 3, very round), which is indicative of readiness to spawn, and noted whether males were stationary (that is, likely to be nest-holding) or not. At each locality we supplied five artificial nest sites (PVC tubes with length 8 cm, diameter 1.4 cm and inner area 32 cm²). These tubes had an inner area very similar to natural mussel nests in the study area (T.A., J.B., E.F. and P. A. Svensson, unpublished data), and could accommodate eggs from about four females (based on a median clutch size of 7.56 cm² observed in the laboratory²⁹). For each sampling occasion we used data on eggs laid (egg area) in the artificial nests ($n = 52$ over the season) to calculate how many additional females would be able to spawn in an average nest. The nests contained more eggs as the season progressed (Table 1), leaving less space for additional clutches, and each male was thus available as a mating option for fewer females. The number of potential further clutches can be viewed as a 'male availability factor'. When estimating the OSR we only included females that were ready to mate (roundness scores of 2 or 3), F_R , and stationary (territorial) males, that is, males that probably had a nest and thus were qualified to mate³⁰. The number of stationary males was multiplied with the availability factor of the respective sampling period, resulting in an estimate of available mating opportunities, from a female's perspective (that is, nest-holding males $\times$ space available for further clutches), M_A . The OSR was then calculated as $OSR = M_A/(M_A + F_R)$.

Competition and courtship

At the same localities and times as mentioned above, we observed focal males ($n = 177$) for 15 min each. About one-third of these males had taken up artificial nest sites, whereas the remaining males were either observed or inferred (from their behaviour) to have natural nests. We recorded all male and female visitors within 30 cm of the focal male and all agonistic and courtship acts occurring within the same distance. Male courtship consists of lateral fin displays and attempts to lead females to the nest; female courtship consists of bending the body so as to maximize display of the colourful belly (sigmoid display)²⁸. Male agonistic behaviour included fin display, chasing and biting. Disputes were normally brief but could sometimes last for the whole observation period. Female agonistic behaviour was subtle, including close following, sigmoid displays directed at conspecifics and occasional chasing.

The propensity of focal males to behave agonistically towards male visitors was calculated as the proportion of visiting males that the focal male behaved aggressively towards. The propensity of females to behave agonistically towards other females (when near to the focal male) was approximated by dividing the number of visiting females behaving agonistically by the maximum number of simultaneously visiting females near

that focal male (excluding replicates without simultaneous female visitors). The propensity to court the opposite sex was calculated for males (proportion of visiting females that were courted by the focal male) and for females (proportion of visiting females that courted the focal male).

We did not include observations of less than 14 min duration (some observations were interrupted, for example by thunderstorms), or observations in which the behaviour of the focal male suggested that he did not have a nest (5 out of 182).

Received 22 November 2003; accepted 13 April 2004; doi:10.1038/nature02562.

1. Darwin, C. *The Descent of Man, and Selection in Relation to Sex* (Murray, London, 1871).
2. Andersson, M. *Sexual Selection* (Princeton Univ. Press, Princeton, 1994).
3. Emlen, S. T. & Oring, L. W. Ecology, sexual selection, and the evolution of mating systems. *Science* **197**, 215–223 (1977).
4. Kvarnemo, C. & Ahnesjö, I. The dynamics of operational sex ratios and competition for mates. *Trends Ecol. Evol.* **11**, 404–408 (1996).
5. Trivers, R. in *Sexual Selection and the Descent of Man 1871–1971* (ed. Campbell, B.) 139–179 (Aldine, Chicago, 1972).
6. Clutton-Brock, T. H. & Vincent, A. C. J. Sexual selection and the potential reproductive rates of males and females. *Nature* **351**, 58–60 (1991).
7. Clutton-Brock, T. H. & Parker, G. A. Potential reproductive rates and the operation of sexual selection. *Q. Rev. Biol.* **67**, 437–456 (1992).
8. Parker, G. A. & Simmons, L. W. Parental investment and the control of sexual selection: Predicting the direction of sexual competition. *Proc. R. Soc. Lond. B* **263**, 315–321 (1996).
9. Kokko, H. & Monaghan, P. Predicting the direction of sexual selection. *Ecol. Lett.* **4**, 159–165 (2001).
10. Wade, M. J. & Shuster, S. M. The evolution of parental care in the context of sexual selection: a critical reassessment of parental investment theory. *Am. Nat.* **160**, 285–292 (2002).
11. Vincent, A., Ahnesjö, I., Berglund, A. & Rosenqvist, G. Pipefishes and seahorses: are they all sex role reversed? *Trends Ecol. Evol.* **7**, 237–241 (1992).
12. Gwynne, D. T. & Simmons, L. W. Experimental reversal of courtship roles in an insect. *Nature* **346**, 172–174 (1990).
13. Gwynne, D. T. Food quality controls sexual selection in mormon crickets by altering male mating investment. *Ecology* **74**, 1406–1413 (1993).
14. Gwynne, D. T., Bailey, W. J. & Annells, A. The sex in short supply for mating varies over small spatial scales in a katydid (*Kawanaphila nartee*, Orthoptera: Tettigoniidae). *Behav. Ecol. Sociobiol.* **42**, 157–162 (1998).
15. Marconato, A., Bisazza, A. & Fabris, N. The cost of parental care and egg cannibalism in the river bullhead, *Cottus gobio* L. (Pisces, Cottidae). *Behav. Ecol. Sociobiol.* **32**, 229–237 (1993).
16. Lindström, K. Effects of resource distribution on sexual selection and the cost of reproduction in sand gobies. *Am. Nat.* **158**, 64–74 (2001).
17. Folstad, I. & Karter, A. J. Parasites, bright males and the immunocompetence handicap. *Am. Nat.* **139**, 603–622 (1992).
18. Macías García, C., Saborío, E. & Berea, C. Does male-biased predation lead to male scarcity in viviparous fish? *J. Fish Biol.* **53**, 104–117 (1998).
19. Kvarnemo, C. Temperature differentially affects male and female reproductive rates in the sand goby—consequences for operational sex-ratio. *Proc. R. Soc. Lond. B* **256**, 151–156 (1994).
20. Ahnesjö, I. Temperature affects male and female potential reproductive rates differently in the sex-role reversed pipefish, *Syngnathus typhle*. *Behav. Ecol.* **6**, 229–233 (1995).
21. Kokita, T. & Nakazono, A. Plasticity in the mating system of the longnose filefish, *Oxymonacanthus longirostris*, in relation to mate availability. *J. Ethol.* **16**, 81–89 (1998).
22. Godwin, J. Phylogenetic and habitat influences on mating system structure in the humbug damselfishes (*Dascyllus*, Pomacentridae). *Bull. Mar. Sci.* **57**, 637–652 (1995).
23. Vincent, A. C. J. & Sadovy, Y. in *Behavioural Ecology and Conservation* (ed. Caro, T.) 209–245 (Oxford Univ. Press, Oxford, 1998).
24. Almada, V. C., Gonçalves, E. J., Oliveira, R. F. & Santos, A. J. Courting females: ecological constraints affect sex-roles in a natural population of the blennioid fish *Salaria pavo*. *Anim. Behav.* **49**, 1125–1127 (1995).
25. Borg, Å. A., Forsgren, E. & Magnhagen, C. Plastic sex roles in the common goby—the effect of nest availability. *Oikos* **98**, 105–115 (2002).
26. Kynard, B. E. Breeding behavior of a lacustrine population of threespine sticklebacks (*Gasterosteus aculeatus* L.). *Behaviour* **67**, 178–207 (1978).
27. Kokko, H. & Johnstone, R. A. Why is mutual mate choice not the norm? Operational sex ratios, sex roles and the evolution of sexually dimorphic and monomorphic signalling. *Phil. Trans. R. Soc. Lond. B* **357**, 319–330 (2002).
28. Amundsen, T. & Forsgren, E. Male mate choice selects for female coloration in a fish. *Proc. Natl Acad. Sci. USA* **98**, 13155–13160 (2001).
29. Bjelvenmark, J. & Forsgren, E. Effects of mate attraction and male–male competition on paternal care in a goby. *Behaviour* **140**, 55–69 (2003).
30. Ahnesjö, I., Kvarnemo, C. & Merilaita, S. Using potential reproductive rates to predict mating competition among individuals qualified to mate. *Behav. Ecol.* **12**, 397–401 (2001).

Acknowledgements We thank A. Berglund, T. Birkhead, R. Hårdling, A. Kazem, H. Kokko, C. Kvarnemo, Å. Lankinen, I. Owens and J. Wright for discussions and comments, and the technical staff at the Kristineberg Marine Research Station for practical help. Our research on the mating dynamics of two-spotted gobies has been supported by the Swedish Natural Sciences Research Council, the Research Council of Norway, the Royal Swedish Academy of Science and the EU Access to Large Scale Facilities program.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to E.F. (elisabet.forsgren@bio.ntnu.no).

Transient cross-reactive immune responses can orchestrate antigenic variation in malaria

Mario Recker¹, Sean Nee², Peter C. Bull³, Sam Kinyanjui⁴, Kevin Marsh^{3,4}, Chris Newbold³ & Sunetra Gupta¹

¹Department of Zoology, University of Oxford, South Parks Road, Oxford OX1 3PS, UK

²Ashworth Laboratories, University of Edinburgh, West Mains Road, Edinburgh EH9 3JT, UK

³Nuffield Department of Clinical Medicine, University of Oxford, John Radcliffe Hospital, Oxford OX3 9DS, UK

⁴Kenya Medical Research Institute, Kilifi, Kenya

The malaria parasite *Plasmodium falciparum* has evolved to prolong its duration of infection by antigenic variation of a major immune target on the surface of the infected red blood cell. This immune evasion strategy depends on the sequential, rather than simultaneous, appearance of immunologically distinct variants. Although the molecular mechanisms by which a single organism switches between variants are known in part^{1–3}, it remains unclear how an entire population of parasites within the host can synchronize expression to avoid rapidly exhausting the variant repertoire. Here we show that short-lived, partially cross-reactive immune responses to parasite-infected erythrocyte surface antigens can produce a cascade of sequentially dominant antigenic variants, each of which is the most immunologically distinct from its preceding types. This model reconciles several previously unexplained and apparently conflicting epidemiological observations by demonstrating that individuals with stronger cross-reactive immune responses can, paradoxically, be more likely to sustain chronic infections. Antigenic

variation has always been seen as an adaptation of the parasite to evade host defence: we show that the coordination necessary for the success of this strategy might be provided by the host.

Chronic infection in *P. falciparum* is believed to be maintained through the sequential variation of highly polymorphic antigens that the parasite inserts into the surface of the infected erythrocyte. These variant surface antigens (VSAs) consist of an immunodominant molecule known as *P. falciparum* erythrocyte membrane protein 1 (PfEMP1) and probably also members of a second family known as rifins¹. Each variant is the target of a specific agglutinating antibody response that has been shown to have a major function in naturally acquired protective immunity^{4–7}. Several field studies suggest that these variant-specific antibodies are long-lasting and lead to a sustained recognition of the infecting parasite isolate^{4–13}. However, it has emerged more recently that *P. falciparum* infection also leads to transient and limited increases in levels of antibodies against VSAs expressed by heterologous isolates^{5,11–14}. Here we unite these observations to suggest a mechanism for the maintenance of antigenic variation that relies solely on the distinction between two types of immune response against VSAs.

We envisage each antigenic variant *i* as comprising, first, a single unique ‘major’ epitope (V_i) that elicits a long lived immune response and, second, several ‘minor’ epitopes that elicit transient immune responses that are not unique to the variant.

Figure 1 presents the results of a simple mathematical model that makes no other assumptions than those described above, showing that differences in duration alone between immune responses to ‘major’ and ‘minor’ epitopes are capable of generating sequential dominance of antigenic variants over an extended period. The panels in Fig. 1 record the time histories of the different antigenic variants that are specified, in addition to their unique high-affinity epitope (V_i), by several possible variants of either two or three minor epitopes. For example, in Fig. 1a there are 64 variants, each containing one of eight possible variants of one minor epitope (for example, variants a, b, c, d, e, f, g and h) and one of eight possible variants (for example l, m, n, p, q, r, s and t) of another minor epitope. All

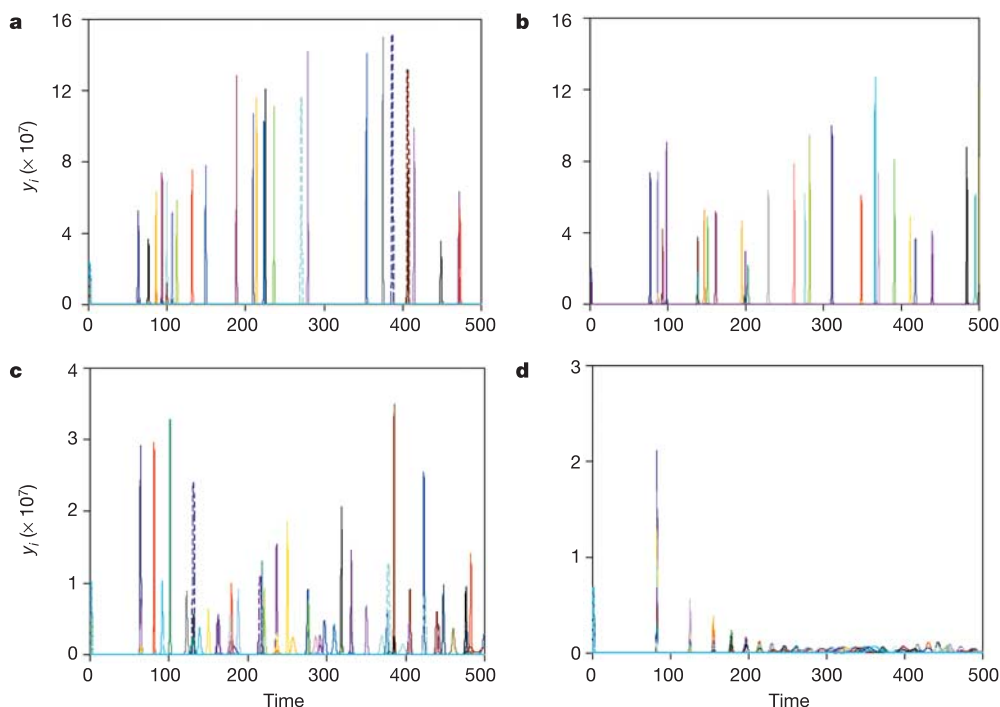


Figure 1 Dynamics of antigenic variants. Time histories are shown of antigenic variants (y_i) specified, in addition to their unique high-affinity epitope, respectively by 8 variants each in 2 minor epitope regions (**a**), 13 variants in 1 minor epitope and 4 in another (**b**), 4 variants each in 3 epitope regions (**c**), and 8 variants each in 2 minor

epitope regions, and one conserved epitope (**d**). Axes are in arbitrary units; parameters used were as follows: $\beta = \beta' = 10^{-4}$; $\alpha = \alpha' = 10^{-3}$; $\mu = 0$; $\mu' = 0.02$; $\phi = 7.5$; initial conditions: all $z, w = 0$, $y_i \in [100, 1,000]$.

variants have the same growth rates and are destroyed at an equal rate by immune responses against minor and major epitopes. The latter are qualitatively identical in every respect other than the duration of the immune response. This simulation illustrates that sequential dominance of antigenic variants will occur provided that the immune response against the shared epitopes decays at a much faster rate than the variant-specific immune response. Figure 1b, c demonstrates that this pattern can be obtained under a variety of combinations of numbers of variants and that it does not rely on any particular choice of number of epitope regions. By contrast, Fig. 1d shows that the inclusion of a strongly immunogenic conserved epitope within the same 64-variant system used in Fig. 1a results in a loss of sequential dominance.

Figure 2b clarifies the underlying mechanism by showing the changes in time in immune responses towards epitopes contained within the antigenic variant marked with an asterisk in Fig. 2a. The latter shows the time course of infection for an example involving three, four and five variants in three epitope regions, as shown in the accompanying diagram. This variant contains a major epitope V_2 and three minor epitopes b, n and x. Short-lived immune responses towards b, n and x can be seen to decline simultaneously at the point in time at which this particular variant emerges, demonstrating that an antigenic variant can be uniquely favoured because of the immune responses generated against the different minor epitopes presented by the preceding sequence of antigenic variants. The development of long-lasting responses specific to epitope V_2 prevents the re-emergence of this variant; the accumulation of such long-lasting specific responses against the entire variant repertoire eventually terminates the infection.

Figure 3 illustrates the generality of this behaviour over a range of variation in the relative efficacy (γ) of the immune responses against the minor epitopes compared with the uniquely variant-specific response with regard to the removal of an antigenic variant. When γ is small, transient responses have no effect and the infection is terminated quickly by the respective long-lasting specific immune responses (Fig. 3a). At the other extreme, when $\gamma = 1$ (that is, the

responses differ only in duration), the infection lasts for a very long period and consists of unique variant peaks, as exemplified by Fig. 1a, b, c for the same system. In between, the dynamics tend from a simultaneous appearance of variants to increasingly sequential patterns with a commensurate increase in the length of the infection, as shown in Fig. 3b, c. The increase in length of infection with γ is shown in more detail in Fig. 4a for an example involving six variants each in two epitope regions; a similar dependence can be shown for parameters that measure induction of immune response rather than effect, and also for an increase in the duration of the 'transient' immune response to minor epitopes. Figure 4a also shows that the intensity of peak parasitaemia declines rapidly as γ increases, because of the disaggregation of variant peaks.

Previous analytical studies have demonstrated that simple differences in growth rates or hierarchies of switch rates cannot sustain sequential dominance of variant peaks^{15–17}. However, it has been shown that optimization of switch rates by natural selection¹⁷ can produce temporal differentiation in the emergence of variants. Here we show that even in the absence of structured differences in switch rates or growth rates, antigenically variable organisms can exhibit a sequential appearance of antigenic variants, provided that there is cross-inhibition between variant antigens that share certain epitopes.

The form of immunological cross-protection in our model is distinct from acquired variant-transcending immunity in that it does not extend across all variants but only involves subsets with common epitopes. Both variant-transcending immunity and immunity against shared epitopes have been shown to promote oscillatory behaviour^{18,19} but not to generate sustained sequential expression. It has also previously been shown that limited levels of cross-protection (for example, extending only to 'nearby' strains) can generate travelling-wave solutions among pathogens exhibiting antigenic drift^{20,21}. This process relies on the accrual of mutations to breach the effects of cross-protection. By contrast, our model is predicated on extensive cross-inhibition and does not explicitly incorporate any switching or mutational phenomena, underscoring the principle that sequential dominance is precipitated by

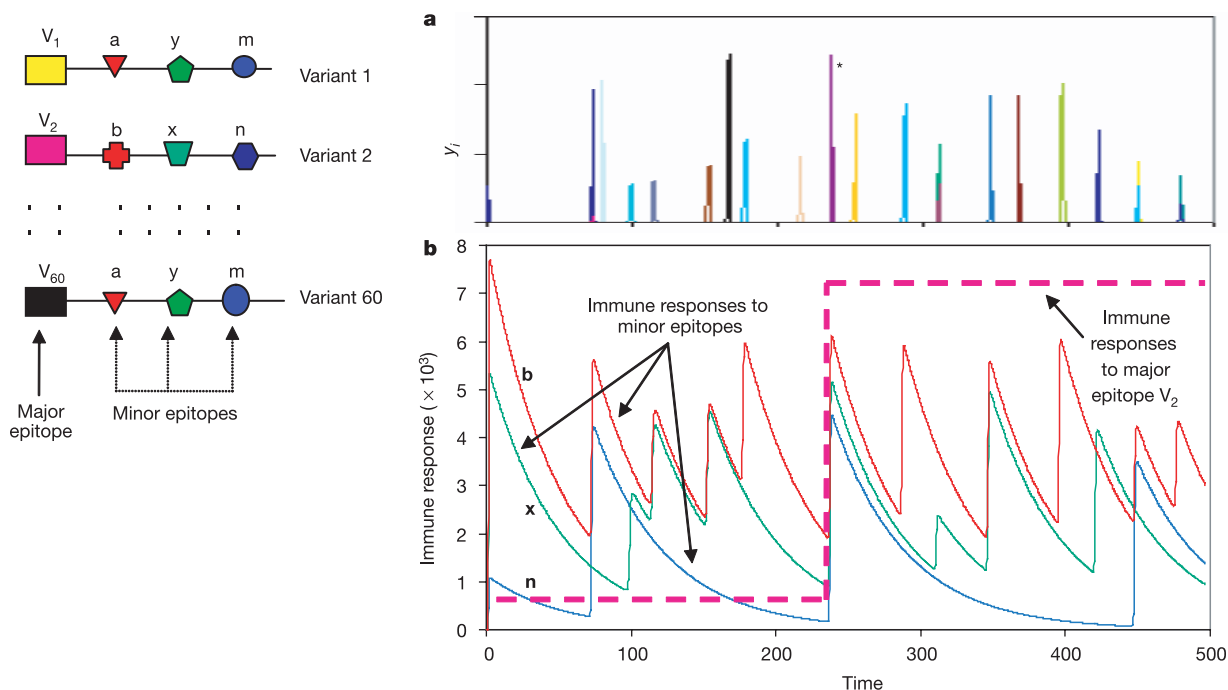


Figure 2 Dynamics of immune response. **a**, Time histories of antigenic variants (y_i) specified, in addition to their unique high-affinity epitope, respectively by three variants in one minor epitope, four in another and five in a third epitope as shown in the accompanying diagram. **b**, Changes with time in immune responses against the minor

variants contained within the antigenic variant marked with an asterisk in **a** as well as the rise in the immune response to its major unique epitope V_2 . All axes are in arbitrary units; parameter values are as in Fig. 1.

immunological cross-protection alone.

Recent attempts to model the within-host dynamics of *P. falciparum*^{22,23} have relied on a very large set of assumptions, many of which remain to be validated experimentally. By contrast, our model contains only two critical assumptions for which there is ample experimental evidence. Several epidemiological studies suggest that VSA contains 'major' unique epitopes that elicit a long-lasting protective immune response^{4–13}, which might serve to structure these into discrete non-overlapping repertoires²⁴. However, there is also evidence now that during the course of infection, individuals are able to react temporarily to a range of heterologous isolates^{5,11,12,14}, suggesting that the isolates share many of the 'minor' epitopes but none of the 'major' epitopes. The kinetics of the immune response in our model are mirrored closely by recent data on temporal changes in levels of VSA antibodies against homologous and heterologous isolates in plasma obtained at monthly intervals from patients in a Ghanaian study¹¹. Our results are also supported by a recent study conducted in Kenya¹³, which reveals that the main bulk of the IgG response to the parasite-infected erythrocyte surface often decays rapidly after treatment, whereas the titre of agglutinating antibodies is maintained. Transient antibody responses against shared epitopes might

also underlie the recent observation that 'partially cross-reactive' antibodies are made after acute malaria infections under conditions of low transmission of malaria in India²⁵. It is also worth noting that 'partially cross-reactive' epitopes have been identified on recombinant PfEMP1 peptides²⁶, although it is not clear whether these correspond to the minor epitopes specified in our model.

Our model results can be used to explain at least two intriguing epidemiological features of *P. falciparum* malaria. The first of these is a long-standing conundrum concerning the age-distribution of *P. falciparum* parasite prevalence. Field studies (for example the Garki Project²⁷) consistently indicate that the duration of infection increases in older children even though the density of parasitaemia and the number of clinical episodes undergo a monotonic decline. This is predicted by our new theoretical framework (see Figs 3 and 4), as a consequence of a shift towards higher γ (that is, the ability to mount a response to minor epitopes) with age due to increased exposure. Furthermore, our model suggests that current infection might increase the duration (but decrease the intensity) of a super-infecting parasite, leading to further chronicity but preventing clinical symptoms. Data on the decline in the odds of clinical attack with multiplicity of infection²⁸ are in strong agreement with this prediction.

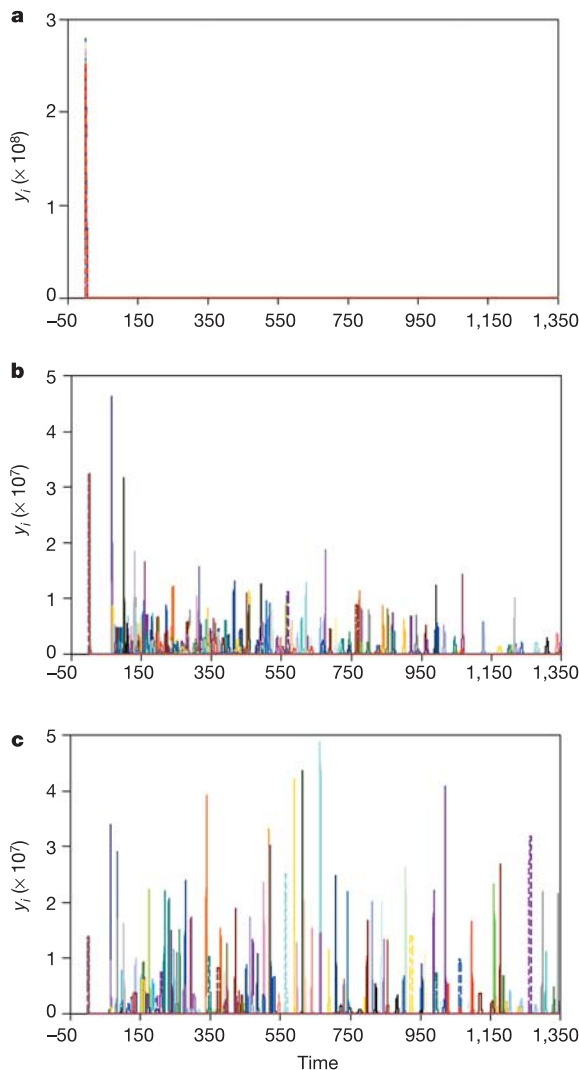


Figure 3 Dependence of dynamics on efficacy of cross-reactive immune response. The dynamics of infection are contrasted for different values of γ (the relative efficacy of the immune response to minor epitopes at either the induction or effector level): **a**, $\gamma = 0$; **b**, $\gamma = 0.25$; **c**, $\gamma = 0.75$ for the same system as in Fig. 1c. The density of parasitaemia bears an inverse relationship to the duration.

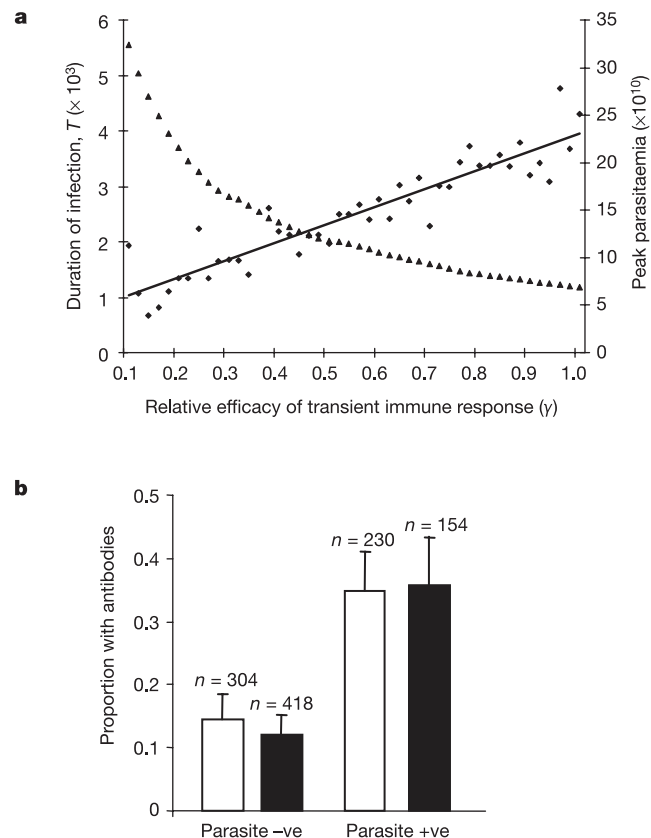


Figure 4 Dependence of duration of infection on efficacy of cross-reactive immune response. **a**, Increase in length of infection, T (arbitrary units), with relative efficacy of the transient immune response, γ , for an example involving six variants each in two epitope regions. Each diamond represents this information for a particular simulation. The corresponding decline in peak density of parasitaemia is shown by the triangles. **b**, Association of heterologous agglutinating antibodies with the presence of asymptomatic parasitaemia at the end of the dry season among 1,106 Kenyan children under two levels of cumulative exposure defined by insecticide-treated bed net usage¹⁴. Filled columns, bed net used; open columns, bed net not used. The odds ratio of having agglutinating antibodies if parasite positive was 3.49 (95% confidence interval 2.56–4.74; $P < 0.01$) after adjustment for age and cumulative exposure. Error bars indicate the upper 95% confidence interval.

This model can also explain the more recent observation that healthy children with microscopically detectable *P. falciparum* infections at the end of the dry season in Kenya are significantly more likely to recognize a heterologous parasite than those without parasites¹⁴ despite similar levels of cumulative exposure, as summarized in Fig. 4b. Within our model framework, this can be ascribed to a difference in individual ability to mount an immune response to minor epitopes and therefore conforms to the relationship between duration of infection and γ shown in Fig. 4a. In other words, individuals who are better able to respond to the minor epitopes (as reflected in their ability to recognize heterologous parasites) are, paradoxically, more likely to sustain a chronic infection.

In summary, by proposing that transient heterologous responses to VSAs have evolved to coordinate the sequential expression of the associated multigene families within the host, we suggest a novel mechanism for antigenic variation in *P. falciparum* that also resolves several conflicting epidemiological observations. It will be interesting to see whether this paradigm extends to other antigenically variable pathogens such as African trypanosomes in which—in addition to structured switch rates and other mechanisms that might initiate an early order of expression^{15,16}—a shared network of minor epitopes between antigenic variants might enable the parasite to exploit the host immune response to achieve chronicity. □

Methods

The dynamics of variant i can be given by

$$dy_i/dt = \phi y_i - \alpha z_i y_i - \alpha' w_i y_i \quad (1)$$

under the assumption that each variant has a net growth rate ϕ and can be destroyed at a rate α by the specific long-lasting immune response (z_i) and at a rate α' by the transient immune responses (w_i) against minor shared epitopes. The relative efficacy (at the effector level) of the transient immune response can be measured as $\gamma = \alpha'/\alpha$.

The dynamics of the specific response, z_i , against strain i can, in its simplest form, be represented as

$$dz_i/dt = \beta y_i - \mu z_i \quad (2)$$

where the proliferation rate β is in direct correlation with the amount of antigen (y_i), and μ is the rate of decay of the immune response. Functions that explicitly incorporate clonal expansion of the relevant immune cells (see Supplementary Information) can be substituted for βy_i in the proliferation term.

The dynamics of the transient cross-reactive immune response against the minor shared epitopes can be represented in the same form as equation (2) with the appropriate parameter changes:

$$dw_i/dt = \beta' \Sigma_j y_j - \mu' w_i \quad (3)$$

where j refers to all variants that share these epitopes with i . The relative efficacy of induction can be measured as β'/β . An equivalence between this measurement and γ ($=\alpha'/\alpha$) can be demonstrated analytically.

Note that this analytical framework intentionally excludes switching between variants as a means of structuring the appearance of antigenic variants.

Received 17 November 2003; accepted 9 March 2004; doi:10.1038/nature02486.

- Newbold, C. Antigenic variation in *Plasmodium falciparum*: mechanisms and consequences. *Curr. Opin. Microbiol.* **2**, 420–425 (1999).
- Scherf, A. *et al.* Antigenic variation in malaria: in situ switching, relaxed and mutually exclusive transcription of var genes during intra-erythrocytic development in *Plasmodium falciparum*. *EMBO J.* **17**, 5418–5426 (1998).
- Deitsch, K. W., Calderwood, M. S. & Welles, T. E. Malaria. Cooperative silencing elements in var genes. *Nature* **412**, 875–876 (2001).
- Bull, P. C. *et al.* Parasite antigens on the infected red cell surface are targets for naturally acquired immunity to malaria. *Nature Med.* **4**, 358–360 (1998).
- Giha, H. A. *et al.* Antibodies to variable *Plasmodium falciparum*-infected erythrocytes surface antigens are associated with protection from novel malaria infections. *Immunol. Lett.* **71**, 117–126 (2000).
- Dodoo, D. *et al.* Antibodies to variant antigens on the surfaces of infected erythrocytes are associated with protection from malaria in Ghanaian children. *Infect. Immun.* **69**, 3713–3718 (2001).
- Tebo, A. E., Kremsner, P. G., Piper, K. P. & Luty, A. J. Low antibody responses to variant surface antigens of *Plasmodium falciparum* are associated with severe malaria and increased susceptibility to malaria attacks in Gabonese children. *Am. J. Trop. Med. Hyg.* **67**, 597–603 (2002).
- Marsh, K. & Howard, R. Antigens induced on erythrocytes by *P. falciparum*: expression of diverse and conserved determinants. *Science* **231**, 150–153 (1986).
- Gupta, S., Trenholme, K., Anderson, R. M. & Day, K. P. Antigenic diversity and the transmission dynamics of *Plasmodium falciparum*. *Science* **263**, 961–963 (1994).
- Bull, P. C., Lowe, B. S., Kortok, M. & Marsh, K. Antibody recognition of *Plasmodium falciparum* erythrocyte surface antigens in Kenya: evidence for rare and prevalent variants. *Infect. Immun.* **67**, 733–739 (1999).
- Ofori, M. F. *et al.* Malaria-induced acquisition of antibodies to *Plasmodium falciparum* variant surface antigens. *Infect. Immun.* **70**, 2982–2988 (2002).
- Giha, H. A. *et al.* Nine-year longitudinal study of antibodies to variant antigens on the surface of

- Plasmodium falciparum*-infected erythrocytes. *Infect. Immun.* **67**, 4092–4098 (1999).
- Kinyanjui, S., Bull, P. C., Newbold, C. I. & Marsh, K. Kinetics of antibody responses to *Plasmodium falciparum*-infected erythrocyte variant surface antigens. *J. Infect. Dis.* **187**, 667–674 (2003).
- Bull, P. C. *et al.* *Plasmodium falciparum* infections are associated with agglutinating antibodies to parasite-infected erythrocyte surface antigens among healthy Kenyan children. *J. Infect. Dis.* **185**, 1688–1691 (2002).
- Kosinski, R. J. Antigenic variation in trypanosomes: a computer analysis of variant order. *Parasitology* **80**, 343–357 (1980).
- Agur, Z., Abiri, D. & Van der Ploeg, L. H. T. Ordered appearance of antigenic variants of African trypanosomes explained in a mathematical model based on a stochastic switch process and immune selection against putative switch intermediates. *Proc. Natl Acad. Sci. USA* **86**, 9626–9630 (1989).
- Frank, S. A. A model for the sequential dominance of antigenic variants in African trypanosome infections. *Proc. R. Soc. Lond. B* **266**, 1397–1401 (1999).
- Antia, R., Nowak, M. A. & Anderson, R. M. Antigenic variation and the within-host dynamics of parasites. *Proc. Natl Acad. Sci. USA* **93**, 985–989 (1996).
- Nowak, M. A. *et al.* Antigenic oscillations and shifting immunodominance in HIV-1 infections. *Nature* **375**, 606–611 (1995).
- Haraguchi, Y. & Sasaki, A. Evolutionary pattern of intra-host pathogen antigenic drift: effect of cross-reactivity in immune response. *Phil. Trans. R. Soc. Lond. B* **352**, 11–20 (1997).
- Gog, J. R. & Grenfell, B. Dynamics and selection of many-strain pathogens. *Proc. Natl Acad. Sci. USA* **99**, 17209–17214 (2002).
- Molineaux, L. *et al.* *Plasmodium falciparum* parasitaemia described by a new mathematical model. *Parasitology* **122**, 379–391 (2001).
- Page-McNichol, S., Gatton, M., Hastings, I. & Saul, A. The *Plasmodium falciparum* var gene switching rate, switching mechanism and patterns of parasite recrudescence described by mathematical modelling. *Parasitology* **124**, 225–235 (2002).
- Gupta, S. *et al.* The maintenance of strain structure in populations of recombining infectious agents. *Nature Med.* **2**, 437–442.
- Chattopadhyay, R. *et al.* *Plasmodium falciparum* infection elicits both variant-specific and cross-reactive antibodies against variant surface antigens. *Infect. Immun.* **71**, 597–604 (2003).
- Gamain, B., Miller, L. H. & Baruch, D. I. The surface variant antigens of *Plasmodium falciparum* contain cross-reactive epitopes. *Proc. Natl Acad. Sci. USA* **98**, 2664–2669 (2001).
- Molineaux, L. & Gramiccia, G. *The Garki Project* (World Health Organisation, Geneva, 1980).
- Beck, H.-P. *et al.* Analysis of multiple *Plasmodium falciparum* infections in Tanzanian children during the trial of the malaria vaccine SPf66. *J. Infect. Dis.* **175**, 921–926 (1997).

Supplementary Information accompanies the paper on www.nature.com/nature.

Acknowledgements We thank A. McLean, G. Rudenko and D. Barry for their valuable comments, and the MRC and The Wellcome Trust for financial support.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to S.G. (sunetra.gupta@zoology.ox.ac.uk).

Positive selection at sites of multiple amino acid replacements since rat–mouse divergence

Georgii A. Bazykin¹, Fyodor A. Kondrashov², Aleksey Y. Ogurtsov³, Shamil Sunyaev⁴ & Alexey S. Kondrashov³

¹Department of Ecology and Evolutionary Biology, Princeton University, Princeton, New Jersey 08544, USA

²Section of Evolution and Ecology, University of California at Davis, Davis, California 95616, USA

³National Center for Biotechnology Information, NIH, Bethesda, Maryland 20894, USA

⁴Division of Genetics, Department of Medicine, Brigham & Women's Hospital, Harvard Medical School, Boston, Massachusetts 02115, USA

New alleles become fixed owing to random drift of nearly neutral mutations or to positive selection of substantially advantageous mutations^{1–3}. After decades of debate, the fraction of fixations driven by selection remains uncertain^{4–9}. Within 9,390 genes, we analysed 28,196 codons at which rat and mouse differ from each other at two nucleotide sites and 1,982 codons with three differences. At codons where rat–mouse divergence involved two non-synonymous substitutions, both of them occurred in

the same lineage, either rat or mouse, in 64% of cases; however, independent substitutions would occur in the same lineage with a probability of only 50%. All three non-synonymous substitutions occurred in the same lineage for 46% of codons, instead of the 25% expected. Furthermore, comparison of 12 pairs of prokaryotic genomes also shows clumping of multiple non-synonymous substitutions in the same lineage. This pattern cannot be explained by correlated mutation or episodes of relaxed negative selection, but instead indicates that positive selection acts at many sites of rapid, successive amino acid replacement.

We aligned 9,390 triplets of orthologous genes from rat, mouse and human. Among the 2,999,920 homologous rat and mouse codons within these genes, 83.30% were identical, and 15.70%, 0.94% and 0.07% differed at one, two and three nucleotide sites, respectively (no-, one-, two- and three-substitution codons). The average evolutionary distances between mouse and rat are 0.22 at synonymous sites (K_s) and 0.04 at non-synonymous sites (K_n), in agreement with previous estimates¹⁰. We assume that at an i -substitution codon, exactly i substitutions occurred after the divergence of rat and mouse lineages from their last common ancestor (the rat–mouse last common ancestor (RMCA)), because non-parsimonious evolutionary paths between such a close pair of species must be rare².

The RMCA codon is revealed exactly by the homologous human codon 'H' if no substitutions occurred on the path connecting these codons. Even after synonymous substitutions, H still reveals the amino acid encoded by the RMCA codon. As the K_s and K_n between human and rat or human and mouse is ~ 0.5 and ~ 0.1 , respectively¹⁰, we expect $\sim 60\%$ of human codons to coincide with the RMCA exactly and $\sim 80\%$ to encode the same amino acid. In fact, among no-substitution codons, H coincides with the codon present in rat 'R' and mouse 'M' in 69% of cases, and encodes the same amino acid, with or without synonymous substitutions, in 90% of cases.

At 71% of one-substitution codons, H coincides with either M or R, and at 74% of one-substitution codons, H encodes the same amino acid as M and/or R. In such cases we assume that H reveals the RMCA codon or, at least, the amino acid it encodes. Otherwise, RMCA remains unknown. We assume that the nucleotide (amino acid) substitution occurred in the rat lineage if H coincides (encodes the same amino acid) with M, and in the mouse lineage if H coincides (encodes the same amino acid) with R (Table 1).

Let us now consider the 28,196 two-substitution codons (Table 2). Among them rat and mouse differ from each other by: two synonymous substitutions (such codons encode either arginine or leucine; for example, TTA versus CTG) at 1,635 codons; one synonymous and one non-synonymous substitution (for example, CCC versus CAT) at 14,935 codons; none or one synonymous substitution and one or two non-synonymous substitutions depending on their order (for example, ACG versus AAT) at 4,417 codons; two non-synonymous substitutions or two synonymous substitutions depending on their order (for example, AGG versus CGT) at 715 codons; and two non-synonymous substitutions (for example, AAA versus AGT) at 6,146 codons. The two substitutions at a codon could occur in the rat lineage (pattern 0), the mouse lineage (pattern 2), or in both lineages (pattern 1). Accordingly, the

RMCA codon would coincide with mouse codon M (pattern 0), rat codon R (pattern 2) or with one of the two intermediate codons I_1 and I_2 (pattern 1; for example, if the rat codon is AAG and the mouse codon is CCG, the intermediate codons are ACG or CAG; for some rat–mouse codon pairs, only one intermediate codon is possible because the other one is a stop codon). When H coincides (or encodes the same amino acid) with M, R, I_1 or I_2 , we assume that it reveals the RMCA codon or, at least, the amino acid it encoded. This was the case for 57% and 62% of codons, respectively. Otherwise, RMCA and the pattern remain unknown.

If the two substitutions were independent (implying that neither of the intermediate codons is a stop codon) and equally common in rat and mouse lineages (Table 1), frequencies of patterns 0, 1 and 2 (P_0 , P_1 and P_2) would be 25%, 50% and 25%, respectively. This is approximately the case when one or both substitutions at a codon were synonymous (Table 2). In contrast, when both substitutions were non-synonymous, we observed a large excess of the frequencies of patterns 0 and 2; that is, of codons where both substitutions occurred within the mouse or the rat lineage. This excess is significant both in comparison with the expected 25:50:25 ratio (chi-square, $P < 0.001$) and in comparison with the pattern in codons with two synonymous substitutions (chi-square, $P < 0.001$). Substantial clumping of non-synonymous substitutions within the same lineage only exists when both substitutions affect the same codon (Supplementary Fig. 1).

Analysis of 1,982 three-substitution codons reveals an even more marked clumping effect (Table 3). For each such codon we need to consider, in addition to R and M, six intermediate codons, three of which differ from R by one substitution (J_1 , J_2 and J_3) and three of which differ from R by two substitutions (K_1 , K_2 and K_3). When the RMCA codon coincides with M, a K codon, a J codon or R the corresponding number of substitutions that occurred in the rat lineage are 3, 2, 1 and 0, respectively (patterns α , β , γ and δ). If the substitutions occurred independently, the ratio of the numbers of codons with patterns α , β , γ and δ would be 1:3:3:1. However, a twofold excess of patterns δ and α is observed, which increases with the contribution of non-synonymous substitutions into rat–mouse divergence. Indeed, this excess is significantly higher when only 0–3 possible paths involve synonymous substitutions than when 4–6 paths involve them (chi-square, $P < 0.001$).

Could this clumping be an artefact? There are two possible sources of error. More than two substitutions may have occurred at a two-substitution codon since rat–mouse divergence. However, if the true number of substitutions at a codon was three, treating it as a two-substitution codon only underestimates the clumping. Indeed, we record pattern 0 (or 2) when H coincides with M (or R), and the presence of an extra substitution on the rat–mouse evolutionary path in such cases implies that three (instead of just two) substitutions occurred within the rat (or mouse) lineage. Clumping can be overestimated only with four or more substitutions at a two-substitution codon, but high degrees of non-parsimony must be very rare for rat and mouse. Furthermore, biased misidentification of RMCA is feasible. We compared data on false excess codons (where evolution on the RMCA–human path may inflate the observed P_0 and P_2) and on false deficit codons (where this evolution may cause underestimation of P_0 and P_2 (Supplementary Information)). The excess of patterns 0 and 2

Table 1 Divergence at codons where rat and mouse differ at one nucleotide site

Parameter	Substitution in rat lineage	Substitution in mouse lineage	RMCA unknown
Synonymous substitution	143,405 (51.2%)	136,475 (48.8%)	88,153 (24.0%)
Non-synonymous substitution			
Nucleotide-level pattern	27,305 (50.4%)	26,868 (49.6%)	48,692 (47.3%)
Amino-acid-level pattern	38,019 (50.3%)	37,583 (49.7%)	27,263 (26.5%)

Frequencies in the first two columns are only within codons where the RMCA is known, either at the nucleotide level or, at least, at the amino acid level (see text).

Table 2 Divergence at codons where rat and mouse differ at two nucleotide sites

Parameter	Both substitutions in rat lineage	One substitution in each lineage	Both substitutions in mouse lineage	RMCA unknown
Two synonymous substitutions	401 (29.0%)	638 (46.1%)	346 (25.0%)	250 (15.3%)
One synonymous and one non-synonymous substitution, neither intermediate codon is a stop	2,449 (27.4%)	4,237 (47.4%)	2,258 (25.3%)	5,900 (39.7%)
One synonymous and one non-synonymous substitution, one intermediate codon is a stop	16 (30.8%)	20 (38.5%)	16 (30.8%)	57 (52.3%)
Two synonymous or two non-synonymous substitutions	130 (28.6%)	197 (43.4%)	127 (28.0%)	261 (36.5%)
None or one synonymous substitution, two or one non-synonymous substitutions	731 (32.3%)	815 (36.0%)	719 (31.7%)	2,152 (48.7%)
Two non-synonymous substitutions, one intermediate codon is a stop, amino-acid-level pattern	78 (38.1%)	57 (27.8%)	70 (34.2%)	125 (37.9%)
Two non-synonymous substitutions, neither intermediate codon is a stop				
Nucleotide-level pattern	875 (30.2%)	1,047 (36.2%)	972 (33.6%)	3,252 (62.9%)
Amino-acid-level pattern at codons:				
All	1,266 (30.3%)	1,543 (36.9%)	1,371 (32.8%)	1,966 (32.0%)
Possible false excess	129 (30.0%)	161 (37.4%)	140 (32.6%)	163 (27.5%)
Possible false deficit	184 (31.5%)	243 (41.5%)	158 (27.0%)	274 (31.9%)
1,3-substitution	120 (28.4%)	163 (38.6%)	139 (32.9%)	193 (31.4%)
CpG-free	822 (30.6%)	971 (36.2%)	891 (33.2%)	1,210 (31.1%)
Convergence-free	165 (25.3%)	254 (39.0%)	233 (35.7%)	304 (31.8%)
Amino-acid-level pattern within regions:				
Very strong conservation	47 (35.6%)	28 (21.2%)	57 (43.2%)	23 (14.8%)
Strong conservation	314 (32.1%)	334 (34.1%)	331 (33.8%)	266 (21.4%)
Moderate conservation	428 (30.7%)	503 (36.1%)	464 (33.3%)	595 (29.9%)
All others	477 (28.5%)	678 (40.5%)	519 (31.0%)	1,082 (39.3%)
Amino-acid-level pattern at genes:				
Low K_n	82 (35.8%)	63 (27.5%)	84 (36.7%)	52 (18.5%)
Medium K_n	341 (30.6%)	400 (35.9%)	372 (33.4%)	404 (26.6%)
High K_n	843 (29.7%)	1,080 (38.1%)	915 (32.2%)	1,510 (34.7%)
Low G+C contents	515 (30.6%)	614 (36.4%)	556 (33.0%)	814 (32.6%)
Medium G+C contents	435 (30.1%)	523 (36.2%)	487 (33.7%)	683 (32.1%)
High G+C contents	316 (30.1%)	406 (38.7%)	328 (31.2%)	469 (30.9%)

Frequencies in the first three columns are only within codons where the RMCA is known. Possible false excess (false deficit) codons are those in which misidentification of RMCA is likely to cause overestimation (underestimation) of the frequency of patterns 0 and 2 (see Supplementary Information). 1,3-substitution codons are those where rat and mouse differ from each other at the first and third nucleotide sites. CpG-free codons are those in which neither of the two possible intermediate states between rat and mouse codons includes CpG context, neither inside the codon nor on its boundary. Convergence-free codons are those where the difference between properties of rat and mouse amino acids is greater than any of the four differences between one of them and one of the two possible intermediate amino acids. Regions with very strong, strong and moderate conservation are those in which the codon under consideration is flanked by gapless rat-mouse-human alignments of length 10 or more each with 10, 8 or 9, and 6 or 7 invariant amino acids, respectively. Genes were split into three equally large bins, according to their rat-mouse K_n , or to their frequency of G and C. Average values of K_n within the bins are 0.006 (low), 0.026 (medium) and 0.081 (high). Average G+C contents within the bins are 0.463 (low), 0.530 (medium) and 0.592 (high).

is only marginally different in these two opposite cases (Table 2; chi-square, $P > 0.1$), ruling out systematic bias in misidentification of RMCA.

In contrast, random misidentification of RMCA will mask the excess of patterns 0 and 2. Even if both substitutions always occur in the same lineage ($P_1 = 0$), we would observe $P_1 \approx$ one-sixth if H randomly deviates from RMCA in one-third of cases (assuming that a deviating H coincides with I_1 or I_2 , instead of M or R, with a probability of 50%). Indeed, P_1 is lowest within genes with low K_n (chi-square, $P < 0.01$) and within conservative regions of all genes (chi-square, $P < 0.001$; Table 2), because misidentification of RMCA is rarer where the divergence is slow.

Data on two-substitution codons within 12 pairs of prokaryotic genomes show that the excess of patterns 0 and 2 for non-synonymous substitution is a universal phenomenon, and is more pronounced when the two sister genomes and the out-group are close to each other (Table 4). Thus, the clumping of non-synonymous substitutions within the same lineage must be real, and misidentification of the ancestral state only obscures it.

Three factors can cause substitutions to occur mostly in the same lineage. First, the corresponding mutations can be correlated. Such correlations can arise if one mutational event simultaneously affects two successive nucleotides; however, such double substitutions are rare¹¹, and the excess of patterns 0 and 2 remains high at codons where rat differs from mouse at the first and the third nucleotides (Table 2). Also, the first substitution can facilitate the second one by increasing the mutation rate. The only known mechanism of such facilitation¹² is the creation of a CpG hypermutable context by the first substitution. However, the excess of patterns 0 and 2 remains high for rat-mouse codon pairs where neither of the intermediate codons involves CpG, either inside the codon or on its boundary. Also, this excess does not depend on the G+C content of a gene (Table 2; chi-square, $P > 0.3$). Peculiarities of the mutation process can hardly be universal, and thus cannot explain a phenomenon observed in many diverse genomes (Table 4). Second, negative selection at a codon can affect one lineage less than the other. This can happen if such selection randomly switches off and on, perhaps due to substitutions at other codons of the same gene

Table 3 Divergence at codons where rat and mouse differ at three nucleotide sites

Number of paths involving synonymous substitutions	All substitutions in rat lineage	Two substitutions in rat and one in mouse	One substitution in rat and two in mouse	All substitutions in mouse lineage	RMCA unknown
Six	114 (20.2%)	160 (28.3%)	167 (29.6%)	124 (21.9%)	378 (40.1%)
Five	38 (24.8%)	45 (29.4%)	38 (24.8%)	32 (20.9%)	113 (42.5%)
Four	18 (17.0%)	30 (28.3%)	31 (29.2%)	27 (25.5%)	103 (49.3%)
Three	32 (32.7%)	20 (20.4%)	16 (16.3%)	30 (30.6%)	100 (50.5%)
Two	13 (24.5%)	13 (24.5%)	11 (20.8%)	16 (30.2%)	46 (46.5%)
One	8 (22.2%)	10 (27.8%)	8 (22.2%)	10 (27.8%)	21 (36.8%)
None	1 (100.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	1 (50.0%)

There are six evolutionary paths between two codons that differ from each other at all three sites, depending on the order in which the substitutions occur. Frequencies in the first four columns are only within codons where the nucleotide-level RMCA is known. Codons for which an intermediate stop codon is possible are not considered.

Table 4 Patterns of divergence between 13 pairs of genomes

Taxon	No. of genes	Fraction of amino acid differences*	All codons with synonymous substitutions†	All codons with non-synonymous substitutions†	Synonymous substitutions at two-substitution codons‡	Non-synonymous substitutions at two-substitution codons‡
Muridae	9,390	4.3% 11.6% 11.6%	143,955:136,963	49,194:48,743	401:638:346 (84.7%) 29.0:46.1:25.0§	1,266:1,543:1,371 (68.0%) 30.3:36.9:32.8§
Bacillus	2,260	3.5% 39.6% 39.5%	18,193:18,065	4,434:4,175	56:128:60 (39.1%) 23.0:52.5:24.6	190:172:151 (42.9%) 37.0:33.5:29.4§
Bordetella	2,878	0.4% 1.3% 1.2%	3,361:2,420	2,066:1,210	1:0:0 (100.0%) 100.0:0.0:0.0	36:3:79 (92.2%) 30.5:2.5:67.0§
Buchnera	421	20.2% 35.2% 35.2%	7,932:7,576	6,187:6,186	67:140:72 (70.6%) 24.0:50.2:25.8	380:484:408 (50.3%) 29.9:38.1:32.1§
Chlamydia	802	11.0% 31.5% 31.4%	22,215:21,908	7,088:6,835	275:502:263 (53.9%) 26.4:48.3:25.3	224:255:296 (42.9%) 28.9:32.9:38.2§
Escherichia	3,590	1.2% 3.4% 3.1%	25,412:21,197	7,151:4,116	255:41:88 (95.5%) 66.4:10.7:22.9§	419:56:164 (91.5%) 65.6:8.8:25.7§
Helicobacter	994	3.0% 42.7% 42.8%	4,500:4,954	1,478:1,606	16:43:19 (24.5%) 20.5:55.1:24.4	50:74:47 (36.7%) 29.2:43.3:27.5
Pseudomonas	2,714	16.9% 26.0% 25.3%	93,293:67,191	45,116:40,321	207:190:81 (65.7%) 43.3:39.8:17.0§	3,422:1,927:3,220 (59.2%) 39.9:22.5:37.6§
Pyrococcus	1,413	13.7% 20.1% 20.4%	43,107:48,275	17,067:18,063	365:698:390 (59.6%) 25.1:48.0:26.8	845:705:1,099 (66.1%) 31.9:26.6:41.5§
Salmonella	2,760	0.8% 12.4% 12.5%	9,230:9,147	2,276:2,466	52:35:37 (80.0%) 41.9:28.2:29.8§	172:37:67 (81.9%) 62.3:13.4:24.3§
Staphylococcus	1,873	0.5% 23.5% 23.5%	2,131:2,163	812:712	5:5:5 (60.0%) 33.3:33.3:33.3	32:17:15 (62.1%) 50.0:26.6:23.4§
Streptococcus	1,197	0.7% 27.3% 27.2%	1,779:1,802	732:638	6:16:9 (60.8%) 19.4:51.6:29.0	12:16:11 (50.0%) 30.8:41.0:28.2
Vibrio	796	0.7% 21.7% 21.7%	2,258:2,339	400:435	6:15:4 (61.0%) 24.0:60.0:16.0	8:9:6 (46.0%) 34.8:39.1:26.1

The following triplets of genomes were analysed, with sister genomes shown in parentheses. Muridae: (*Rattus norvegicus*, *Mus musculus*), *Homo sapiens*; *Bacillus*: (*B. cereus* ATCC 14579, *B. anthracis* strain Ames), *B. subtilis subtilis* strain 168; *Bordetella*: (*B. parapertussis*, *B. bronchiseptica* RB50), *B. pertussis* Tohama I; *Buchnera*: (*B. aphidicola* strain APS, *B. aphidicola* strain Sg), *B. aphidicola* strain Bp; *Chlamydia*: (*C. trachomatis*, *C. muridarum*), *C. caviae* GPIC; *Escherichia*: (*E. coli* O157:H7, *E. coli* K12), *E. coli* CFT073; *Helicobacter*: (*H. pylori* 26695, *H. pylori* J99), *H. hepaticus* ATCC 51449; *Pseudomonas*: (*P. syringae* pv. tomato strain DC3000, *P. putida* KT2440), *P. aeruginosa* PAO1; *Pyrococcus*: (*P. horikoshii*, *P. abyssi*), *P. furiosus* DSM 3638; *Salmonella*: (*S. typhimurium* LT2, *S. enterica enterica* serovar Typhi Ty2), *Shigella flexneri* 2a strain 301; *Staphylococcus*: (*S. aureus aureus* Mu50, *S. aureus aureus* MW2), *S. epidermidis* ATCC 12228; *Streptococcus*: (*S. pyogenes* M1 GAS, *S. pyogenes* MGAS315), *S. agalactiae* NEM316; *Vibrio*: (*V. vulnificus* YJ016, *V. vulnificus* CMCP6), *V. parahaemolyticus* RIMD 2210633.

*Fraction of mismatches in alignments of orthologous proteins between sister genome 1 and sister genome 2, sister genome 1 and the out-group, and sister genome 2 and the out-group, respectively.

†Total number of codons at which at least one synonymous, or at least one non-synonymous, substitution occurred in the first and the second sister lineage after their divergence.

‡Data on two-substitution codons where both substitutions were either synonymous or non-synonymous. The top line shows the number of codons where both substitutions occurred in sister lineage 1, one substitution occurred in each lineage, and both substitutions occurred in lineage 2, respectively. The fraction of codons for which the out-group reveals the last common ancestor is given in parentheses. The bottom line shows the frequencies of the corresponding patterns of substitution among those codons for which the common ancestor is known.

§Significance of deviation from 25:50:25 ($P < 0.01$).

(covarion model^{13,14}) or of other genes. However, quantitative analysis of fluctuating negative selection shows that it is unlikely to generate the observed excess of patterns 0 and 2 (Supplementary Fig. 2).

The third (and the only remaining) explanation is positive selection. One possibility is that at a two-substitution codon, the first substitution was slightly deleterious and only the successive, compensatory substitution was advantageous. Such situations can be expected when the amino acid encoded after two substitutions is more similar to the original one than the amino acid encoded after the first substitution. We tested this possibility by only considering those two-substitution codons where the difference in properties (according to Miyata matrix¹⁵) between amino acids encoded by M and R is greater than any of the differences between amino acids encoded by I₁ and M, I₁ and R, I₂ and M, and I₂ and R. The excess of patterns 0 and 2 at such convergence-free codons was essentially the same as at other codons (Table 2; chi-square, $P > 0.2$).

Thus, positive selection usually favours every successive non-synonymous substitution at a codon. Indeed, when selection favours a new amino acid, there is no reason why this amino acid should always be reachable by just a single nucleotide substitution. This positive selection is probably different between the rat and mouse lineages, as identical selection would lead to homoplasy and their parallel evolution. However, the overall rate of evolution of a protein is principally determined not by positive selection but by selective constraint, because the combined excess of patterns 0 and 2 is the highest for slowly evolving proteins (Table 2).

A rather small (3–4%) excess of patterns 0 and 2 exists even when one (chi-square, $P < 0.001$) or both (chi-square, $P < 0.005$) substitutions at a two-substitution codon were synonymous (Table 2). Such excesses are also present in several pairs of bacterial genomes

(Table 4). This is consistent with very weak selection on silent sites in mammalian^{16,17} and in bacterial² genomes.

To generate the observed excesses of two-substitution codons with patterns 0 and 2, and of three-substitution codons with patterns α and δ, positive selection must act on a substantial fraction of non-synonymous substitutions at these codons. As there were over 2,500 of such substitutions in 9,390 genes, and the total number of amino acid replacements between mouse and rat is approximately 25 per protein (Table 4), the fraction of replacements driven by positive selection is at least about 0.5%. A tendency of neighbouring amino acid replacements to occur within the same lineage, which slowly declines when the distance between them increases (Supplementary Fig. 1), suggests that positive selection also affects some one-substitution codons. Thus, the overall contribution of positive selection may be considerably higher^{4–9,18}. □

Methods

Analysis of alignments

Human, mouse and rat coding sequences were obtained from version 32 of the human genome¹⁹, version 30 of the mouse genome²⁰ and version 2 of the rat genome²¹ from the National Center for Biotechnology Information (NCBI). Orthologues were identified according to the two-directional best-hit approach tailored for three species²² using protein BLAST²³. Alignments of the amino acid sequences for each of the orthologue triplets were made using ClustalW²⁴ and reverse transcribed to get the nucleotide triple alignments. K_s and K_n were estimated using pairwise nucleotide alignments taken from the triple alignments for mouse–rat, mouse–human and rat–human pairs using the codeml program of the PAML package²⁵. To eliminate erroneous alignments and non-orthologous gene triplets, those triplets that showed $K_n > 0.2$ and/or $K_s > 0.45$ in the mouse–rat comparison were removed from the analysis. To eliminate erroneous regions of alignments, which may originate from errors in genome assemblies or annotations, we only analysed those codons that were flanked by gapless alignments of length ten or more with at least five amino-acid matches between rat and mouse sequences (on each side) and at least three matches between human and rat and/or mouse sequences. Difference between a pair of amino acids was taken as the corresponding term from Miyata matrix of

amino-acid pair distances¹⁵. Subdivision of rat–mouse–human codon triplets into classes represented in Tables 1–3 easily follows from the genetic code table. All suitable triplets of bacterial genomes were obtained from the NCBI Entrez database and processed analogously. All alignments are available at <http://ftp.ncbi.nih.gov/pub/kondrashov/RatMouse/>.

Fluctuating negative selection

We assumed that negative selection at a codon switches off and on at random moments. The expected waiting times (in units of time since rat–mouse divergence) for off to on and on to off switches are T and bT , respectively. Thus, negative selection is present with a probability of $b/(1+b)$. If the total duration of episodes of absent negative selection in rat and mouse lineages were f_r and f_m , respectively, at a two-substitution codon $P_0 = r^2$, $P_1 = 2r(1-r)$ and $P_2 = (1-r)^2$, where $r = f_r/(f_r + f_m)$. This model was studied by Monte–Carlo simulations. For each pair of values of T and b , we performed 1,000,000 runs. Initially, negative selection in both rat and mouse lineages was off with probability $1/(1+b)$ and on with probability $b/(1+b)$. Then, switches of negative selection and accumulation of substitutions in the two lineages occurred independently. For each run, we calculated the probability that at a codon exactly two substitutions took place in rat and mouse lineages, assuming that substitutions occur independently, only when negative selection is off, with the instant rate 0.2 (the value of rat–mouse K_a). After this, the probability of pattern 1 was calculated within two-substitution codons.

Received 5 January; accepted 26 April 2004; doi:10.1038/nature02601.

- Kimura, M. *The Neutral Theory of Molecular Evolution* (Cambridge Univ. Press, Cambridge, 1983).
- Gillespie, J. H. *The Causes of Molecular Evolution* (Oxford Univ. Press, Oxford, 1991).
- Ohta, T. Near-neutrality in evolution of genes and gene regulation. *Proc. Natl Acad. Sci. USA* **99**, 16134–16137 (2002).
- Smith, N. G. C. & Eyre-Walker, A. Adaptive protein evolution in *Drosophila*. *Nature* **415**, 1022–1024 (2002).
- Fay, J. C., Wyckoff, G. J. & Wu, C.-I. Testing the neutral theory of molecular evolution with genomic data from *Drosophila*. *Nature* **415**, 1024–1026 (2002).
- Bustamante, C. D. *et al.* The cost of inbreeding in *Arabidopsis*. *Nature* **416**, 531–534 (2002).
- Eyre-Walker, A. Changing effective population size and the McDonald–Kreitman test. *Genetics* **162**, 2017–2024 (2002).
- Yang, Z. H. Inference of selection from multiple species alignments. *Curr. Opin. Genet. Dev.* **12**, 688–694 (2002).
- Anisimova, M., Nielsen, R. & Yang, Z. H. Effect of recombination on the accuracy of the likelihood method for detecting positive selection at amino acid sites. *Genetics* **164**, 1229–1236 (2003).
- Makalowski, W. & Boguski, M. S. Evolutionary parameters of the transcribed mammalian genome: an analysis of 2,820 orthologous rodent and human sequences. *Proc. Natl Acad. Sci. USA* **95**, 9407–9412 (1998).
- Kondrashov, A. S. Direct estimates of human per nucleotide mutation rates at 20 loci causing Mendelian diseases. *Hum. Mutat.* **21**, 12–27 (2003).
- Nachman, M. W. & Crowell, S. L. Estimate of the mutation rate per nucleotide in humans. *Genetics* **156**, 297–304 (2000).
- Fitch, W. M. & Markowitz, E. An improved method for determining codon variability in a gene and its application to the rate of fixation of mutations in evolution. *Biochem. Genet.* **4**, 579–593 (1970).
- Huelsenbeck, J. P. Testing a covariate model of DNA substitution. *Mol. Biol. Evol.* **19**, 698–707 (2002).
- Miyata, T., Miyazawa, S. & Yasunaga, T. Two types of amino acid substitutions in protein evolution. *J. Mol. Evol.* **12**, 219–236 (1979).
- Bustamante, C. D., Nielsen, R. & Hartl, D. L. A maximum likelihood method for analyzing pseudogene evolution: implications for silent site evolution in humans and rodents. *Mol. Biol. Evol.* **19**, 110–127 (2002).
- Hellman, I. *et al.* Selection on human genes as revealed by comparisons to chimpanzee cDNA. *Genome Res.* **13**, 831–837 (2003).
- Wilbur, W. J. On the PAM matrix model of protein evolution. *Mol. Biol. Evol.* **2**, 434–447 (1985).
- International Human Genome Sequencing Consortium. Initial sequencing and analysis of the human genome. *Nature* **409**, 860–921 (2001).
- Mouse Genome Sequencing Consortium. Initial sequencing and comparative analysis of the mouse genome. *Nature* **420**, 520–562 (2002).
- Rat Genome Sequencing Project Consortium. Genome sequence of the Brown Norway rat yields insights into mammalian evolution. *Nature* **428**, 493–521 (2004).
- Tatusov, R. L., Koonin, E. V. & Lipman, D. J. A genomic perspective on protein families. *Science* **278**, 631–637 (1997).
- Altschul, S. F. *et al.* Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* **25**, 3389–3402 (1997).
- Thompson, J. D., Higgins, D. G. & Gibson, T. J. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res.* **22**, 4673–4680 (1994).
- Yang, Z. H. PAML: a program package for phylogenetic analysis by maximum likelihood. *Comput. Appl. Biosci.* **13**, 555–556 (1997).

Supplementary Information accompanies the paper on www.nature.com/nature.

Acknowledgements We thank N. Bierne for a number of suggestions. G.A.B. was supported by a BWF graduate fellowship. S.S. was supported by Genome Canada Foundation.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to G.A.B. (gbazykin@princeton.edu).

Drosophila dFOXO controls lifespan and regulates insulin signalling in brain and fat body

Dae Sung Hwangbo, Boris Gersham, Meng-Ping Tu, Michael Palmer & Marc Tatar

Division of Biology and Medicine, Box G-W, Brown University, Providence, Rhode Island 02912, USA

In *Drosophila melanogaster*, ageing is slowed when insulin-like signalling is reduced: life expectancy is extended by more than 50% when the insulin-like receptor (*InR*) or its receptor substrate (*chico*) are mutated, or when insulin-producing cells are ablated^{1–3}. But we have yet to resolve when insulin affects ageing, or whether insulin signals regulate ageing directly or indirectly through secondary hormones. *Caenorhabditis elegans* lifespan is also extended when insulin signalling is inhibited in certain tissues, or when repressed in adult worms^{4,5}, and this requires the forkhead transcription factor (FOXO) encoded by *daf-16* (ref. 6). The *D. melanogaster* insulin-like receptor mediates phosphorylation of dFOXO, the equivalent of nematode *daf-16* and mammalian FOXO3a^{7,8}. We demonstrate here that dFOXO regulates *D. melanogaster* ageing when activated in the adult pericerebral fat body. We further show that this limited activation of dFOXO reduces expression of the *Drosophila* insulin-like peptide *dilp-2* synthesized in neurons, and represses endogenous insulin-dependent signalling in peripheral fat body. These findings suggest that autonomous and non-autonomous roles of insulin signalling combine to control ageing.

To investigate whether activated dFOXO affects ageing in *D. melanogaster* we conditionally expressed dFOXO in specific adult tissues. Without ligand binding at the insulin-like receptor, dFOXO remains unphosphorylated and is transported to the nucleus where it promotes factors that retard cell growth and proliferation^{7,8}. We transformed *D. melanogaster* with UAS-constructs, containing either a wild-type full-length complementary DNA of dFOXO (UAS–dFOXO) or dFOXO with the three protein kinase B (PKB) phosphorylation sites mutated to permit insulin-insensitive nuclear transport (UAS–dFOXO-TM). Expression of these constructs in the eye disc reduced growth (Supplementary Fig. S1), as has previously been reported for independent transformants of UAS–dFOXO and for a phosphorylation-site mutant of human FOXO3a^{7,8}. The constitutive expression of UAS–dFOXO or UAS–dFOXO-TM killed larvae when promoted from *actin*-GAL4, or when expressed from fat body (*adh*-GAL4) or neurons (*ELAV*-GAL4) (Supplementary Table S1). Therefore, conditional expression of dFOXO is required to bypass developmental lethality as well as to study its impact on ageing exclusively in the adult stage.

We used the mifepristone inducible-GAL4 system (annotated P{Switch}⁹ and GeneSwitch¹⁰) to drive the expression of UAS-constructs in defined adult tissues. Ingested mifepristone strongly induced reporter expression at all ages (Supplementary Fig. S1), and the compound alone had no effect on adult survival (Supplementary Fig. S2). Adult survival was not improved when UAS–dFOXO-TM was induced by a pan-neuronal driver (*ELAV*-GeneSwitch), or in glial cells (P{Switch} MB221) or neurolemma (P{Switch} S₁13) (Fig. 1; Supplementary Table S2). Thus, broadly activated dFOXO in neuron-associated cells is not sufficient to slow ageing; however, it may do so if expressed in subsets of cells within these tissues. Similarly, expression of UAS–dFOXO-TM or UAS–dFOXO did not affect survival when induced with the P{Switch} strain S₁106, an efficient promoter in the fat body⁹. In contrast, survival was significantly increased in both sexes when dFOXO was induced

with the P{Switch} strain S₁₃₂, which is also expressed in fat body⁹. Multiple independent inserts of UAS-*dFOXO*-TM and of wild-type UAS-*dFOXO* increased median lifespan by as much as 35% when induced with 25 $\mu\text{g ml}^{-1}$ mifepristone and 56% when induced by 50 $\mu\text{g ml}^{-1}$ mifepristone; averaged across trials, lifespan was increased by 15.5% in males and 19.4% in females (Fig. 1; Supplementary Table S2). Because the phosphatase and tensin homologue protein (PTEN) antagonizes phosphatidylinositol-3-OH kinase activity, which promotes nuclear localization of endogenous *dFOXO* and inhibits TOR function, we induced UAS-*dPTEN* with S₁₁₀₆ and S₁₃₂. Survival was unaffected when *PTEN* was expressed from S₁₁₀₆ but increased by about 20% when expressed from S₁₃₂

(Fig. 1; Supplementary Table S2). Together, these data demonstrate that *dFOXO* activated in a specific tissue can regulate lifespan in adult *D. melanogaster*.

To understand how S₁₃₂ but not S₁₁₀₆ can improve lifespan we compared their patterns of expression in adult tissue (Fig. 2). S₁₁₀₆ was expressed in fat body of the thorax, abdomen and occasionally in the cavity surrounding the mouthparts. In contrast, S₁₃₂ was found expressed in fat body of the head and not in the abdomen or thorax (Fig. 2; Supplementary Fig. S5); S₁₃₂ uniquely appears in the pericerebral fat body located above the brain⁹. We mapped the insertion site of S₁₃₂ to the first intron of *bunched*. Although *bunched* was identified in egg follicles¹¹, S₁₃₂ does not express in

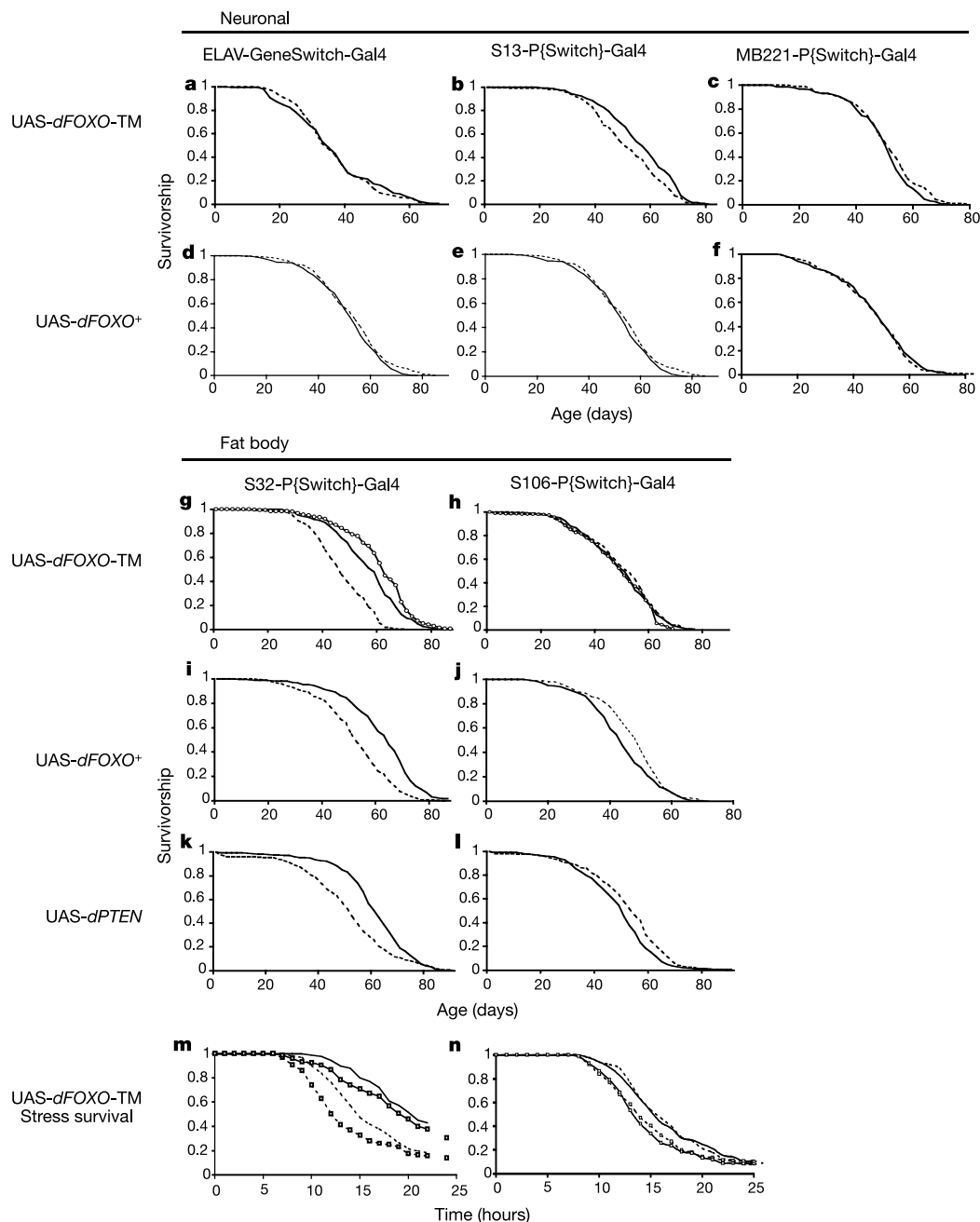


Figure 1 Survivorship and stress resistance in adults with tissue-specific expression of *dFOXO* or *dPTEN*. Each panel represents a genotype defined by the mifepristone-Gal4 driver (column) and the transgene *dFOXO* or *dPTEN* (row). **a–l**, Experiments performed on male flies. Control (dashed line), 25 μg mifepristone (solid line) and 50 μg mifepristone

(circle-marker line; in **g** and **h**). The complete survival data for females and males, and for independent trials and inserts can be found in Supplementary Table S2. **m** and **n**, Stress survival of males (no marker) and females (square marker). Five-day-old adults were fed yeast paste without (broken line) or with (solid line) 25 μg mifepristone.

these cells (Fig. 2e), perhaps because it is inserted near two intronic open reading frames. We conclude that specifically activated dFOXO in the adult head fat body is sufficient to slow ageing.

Surprisingly, although *UAS-dFOXO-TM* induced by *S₁32* increased longevity it did not affect fecundity (Supplementary

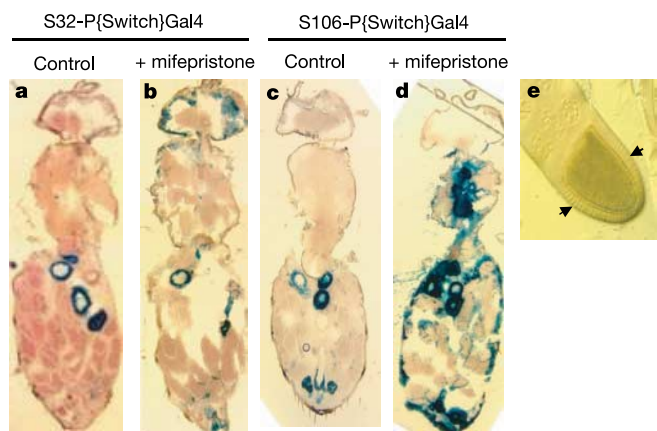


Figure 2 Tissue distribution of fat body drivers *S₁32* and *S₁106*. Longitudinal sections of females at 4 weeks, reported by β -gal staining of *UAS-lacZ*. **a–d**, Patterns of *S₁32* (**a, b**) and patterns of *S₁106* (**c, d**) without (control) (**a, c**) and with mifepristone (**b, d**). **e**, Egg chamber from *S₁32/P[UAS-lacZ]* female with mifepristone; expression is absent in follicle cells (arrows).

Fig. S2). On the other hand, when we challenged adults with an acute oxidative stress agent (paraquat), survival was improved when *UAS-dFOXO-TM* was induced by *S₁32* (in the head fat body) but not by *S₁106* (Fig. 1), in agreement with long-lived insulin signalling mutants of *C. elegans* and *D. melanogaster* that are often stress resistant¹². Similarly, lipids are frequently elevated in *C. elegans* and *D. melanogaster* insulin-signalling mutants^{13,14} and as anticipated, when *dFOXO-TM* was expressed in the head fat body, lipid aggregates appeared in this tissue (Fig. 3h). Remarkably, in the same animals, lipids also accumulated in the peripheral fat tissue even though this construct was not expressed outside the head (Fig. 3p).

To understand how *dFOXO-TM* that is expressed exclusively in the head fat body can regulate integrated physiological traits such as ageing, stress resistance and lipid metabolism, we followed the cellular location of the dFOXO protein in the head and peripheral fat body (Fig. 3). Without transgene induction, endogenous dFOXO was distributed throughout the cytoplasm in all cells. On expression of *UAS-dFOXO-TM*, antibody-labelled dFOXO was increased in both the cytoplasm and nuclei of the targeted tissue. Notably, dFOXO-TM induced by *S₁32* in head fat body also increased endogenous dFOXO nuclear localization of peripheral fat body, in agreement with the pattern of lipid accumulation. In contrast, *dFOXO-TM* expressed in peripheral fat body did not affect endogenous dFOXO in the head (Supplementary Fig. S3). As endogenous dFOXO in peripheral tissue can become localized in the nucleus in response to decreased insulin signalling, these results suggest that dFOXO that is activated in the head fat body retards systemic levels of the insulin ligand.

Cells in the *pars intercebralis* of the adult brain synthesize insulin

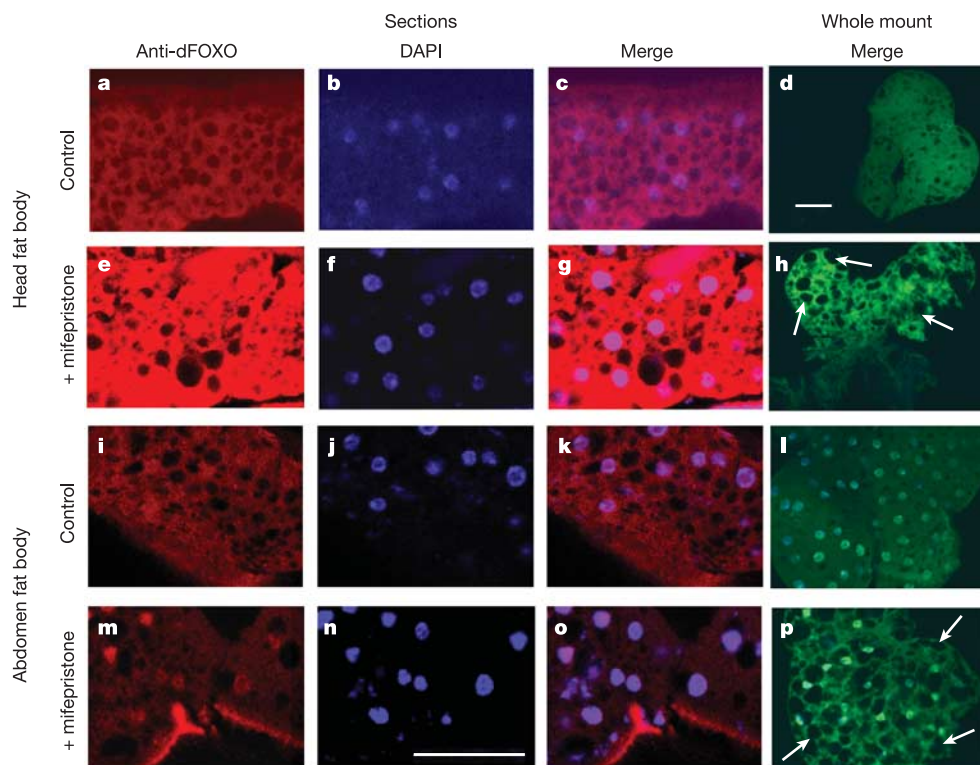


Figure 3 Cellular localization of dFOXO in head and abdominal fat body. **a–d**, Endogenous dFOXO in head fat body has weak nuclear localization (control). **e–h**, The same tissue when exogenous *dFOXO-TM* is expressed by feeding mifepristone; dFOXO is saturated in cytoplasm and increased in nuclei. **i–l**, Endogenous dFOXO in abdominal fat body with weak nuclear localization. **m–p**, shows the decrease in the cytoplasmic distribution of

endogenous dFOXO and an increase in its nuclear localization when exogenous *dFOXO-TM* is expressed in the head fat body. Whole tissue (**h** and **p**) shows dFOXO nuclear localization and increase in lipid aggregates (arrows). Scale bar is 50 μ m on sections and 40 μ m on whole mount tissue.

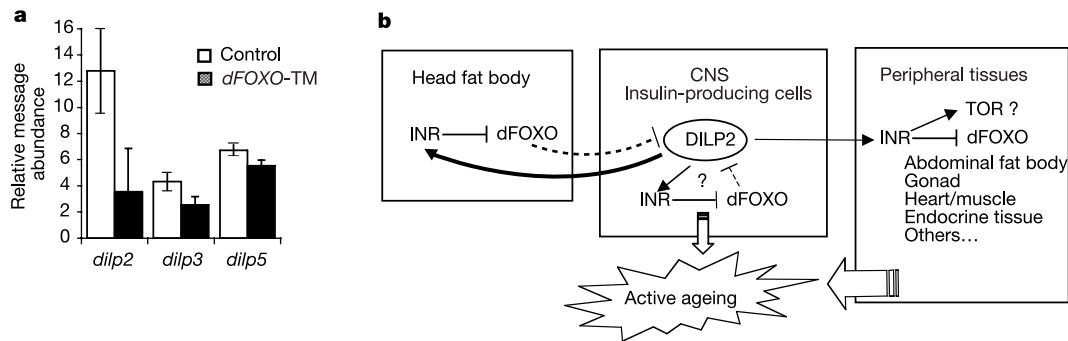


Figure 4 The abundance and systemic impact of neuronal *dilp* mRNA in response to dFOXO expressed in head fat body. **a**, Relative message based on real-time quantitative PCR (with standard error). *dilp2* shows 73% decrease in message upon head fat body expression of dFOXO-TM (nested ANOVA, $p = 0.0014$); no significant change for *dilp3* ($p = 0.27$) or *dilp5* ($p = 0.44$). **b**, Model of ageing regulated by insulin signals. The head fat body and CNS participate in a non-autonomous regulatory circuit with insulin acting upon the head fat body (heavy solid arrow) and the fat body controlling an unknown secondary signal through dFOXO that suppresses *dilp2* transcription (heavy broken line). Within the CNS, insulin signals may participate in a

paracrine circuit to coordinate ligand synthesis of the multiple insulin producing cells. Ligand of *dilp2* released into the circulation (solid arrow) influences insulin-dependent, tissue-autonomous somatic senescence, illustrated by the case of cardiac ageing (R. J. Wessells, E. Fitzgerald, J. R. Cypser, M.T. & R. Bodmer, unpublished data). Autonomous effects of insulin at tissue may exclusively act through dFOXO or include action through TOR^{25,26}. Ligand of *dilp2* may also influence the production of secondary ageing regulatory signals from endocrine tissues (juvenile hormone) or from gonads (ecdysone).

peptides^{15,16}. To test whether activated dFOXO in the head fat body influences insulin production we measured messenger RNA levels of the seven *Drosophila insulin-like peptides* (*dilp*)¹⁷. Complementary DNA was prepared from the heads of S₁32 /UAS-dFOXO-TM adults fed mifepristone or treated as controls. In a preliminary screen, multiple independent samples were specifically analysed for *dilp* message abundance using microarrays: *dilp-2* alone was reduced in response to activated dFOXO in the head fat body (Supplementary Fig. S4). We used these samples to perform quantitative polymerase chain reaction with reverse transcription (RT-PCR) to determine robustly the relative abundance of *dilp* message originating in the adult brain: the *dilp-2* message decreased nearly threefold whereas *dilp-3* and *dilp-5* were unchanged (Fig. 4a). Therefore, insulin signalling within the head fat body influences transcription of one specific *dilp* of the neuronal insulin-producing cells.

Studies across model systems have established that insulin-like signalling can control lifespan non-autonomously from a limited set of cells or a specific tissue. In *C. elegans*, these cells may occur primarily in the intestine and secondarily in neurons^{5,18}. In mice, a disrupted insulin receptor in the adipose tissue across all life stages altered adult adipose morphology, decreased fasted insulin levels and modestly increased adult survival^{19,20}. Here we show with the fly that activated dFOXO in the head fat body is sufficient to increase both male and female lifespan, to increase resistance to oxidative challenge and to alter whole-animal lipid metabolism. Therefore, in *D. melanogaster* systemic secondary signals must function downstream of dFOXO, activated in the head fat body. Both juvenile hormone and 20-hydroxyecdysone are reduced in *D. melanogaster* mutants of *InR*, and both these hormones have the potential to regulate lifespan^{3,21}. Candidates for secondary hormonal signals have yet to be identified in *C. elegans* but these may involve sterols because *daf-9*, a cytochrome P450 related to mammalian steroidogenic hydroxylases, functions downstream of *daf-2* but upstream of *daf-12*, which encodes a putative nuclear hormone receptor^{22,23}.

Our data suggest that an insulin peptide itself may function as one secondary messenger of insulin-regulated ageing in *D. melanogaster* (Fig. 4b). A similar model is emerging for *C. elegans*: the peptide encoded by *ins-7* accelerates ageing and is a systemic agonist of the *daf-2* encoded receptor, whereas functional DAF-16 in the

intestine non-autonomously activates DAF-16 in distant tissues and is sufficient to increase lifespan^{5,24}. We now find that *dilp-2* is uniquely reduced when dFOXO is activated in the head fat body of *D. melanogaster*; this transcriptional change will decrease the amount of circulating insulin peptide released from insulin-producing neurons. Notably, ablation of neuronal cells expressing *dilp-2* is sufficient to retard demographic ageing and to retard functional decline of the adult heart (R. J. Wessells, E. Fitzgerald, J. R. Cypser, M.T. & R. Bodmer, unpublished data). Ageing of the heart, however, can also be delayed when insulin signalling is inhibited exclusively within the cardiac tissue itself, including heart-specific expression of dFOXO-TM. If senescence of tissue and systems throughout the adult is regulated autonomously by insulin, decreased circulation of this peptide downstream of regulatory insulin action within the head fat body could extend lifespan by reducing the mortality risk associated with degeneration of the soma as a whole. □

Methods

Drosophila strains

ELAV-GeneSwitch-Gal4¹⁰ was provided by T. Osterwalder. The P[Switch]-Gal4 enhancer trap strains⁹ and their expression patterns were provided by G. Roman. UAS-*PTE*n was *yw* *hs-flp*¹²²; UAS-d*PTE*n (III), provided by B. Edgar. With pUAST plasmids containing dFOXO (described in ref. 8; provided by A. Brunet) we conducted P-element-mediated germline transformation in an *yw* background and recovered multiple homozygous viable strains with a single insert of UAS-dFOXO-TM (construct with PKB phosphorylation mutations T44A, S190A and S259A) or of UAS-dFOXO⁺ (construct of wild-type cDNA). Animals for all described assays were generated from crosses of virgin females from the specified Gal4 strain with males of each dFOXO or d*PTE*n strain.

Survivorship

Adult survivorship was estimated by the extinct cohort method, with data combined across 3 to 5 replicate demography cages per treatment for each genotype. Each demography cage was initiated with 150 newly eclosed adults, mixed sex; dead flies were removed and scored every two days, at which time fresh food was provided in a vial with 3 ml of cornmeal-sugar-agar media and 0.5 ml of yeast paste (on the vial wall). Yeast paste contained live yeast and water at ratio of 2.5:1 plus 0.5% acetic acid and 1% ethanol (with or without mifepristone). Mifepristone was at 25 µg ml⁻¹ in the yeast paste, unless otherwise stated. Stress survival was measured over 24 h when adults were exposed to paraquat (30 µM in solution with 1% sucrose and 1% ethanol, with or without mifepristone).

Immuno-histology

Tissues for cryosection and for whole mount were isolated from the pericerebral fat body of the head, and from peripheral fat body from the posterior abdomen of 4-day-old females. Tissues were incubated with the anti-dFOXO antibody (ref. 7; provided by O. Puig), and then labelled with rhodamine-conjugated (sections) or Alexa-conjugated

(whole mount) anti-rabbit IgG secondary antibody, followed by staining with DAPI. For sections, all images were captured using a standardized exposure.

Quantitative PCR

The cDNA was prepared from 100 heads per sample of 10-day-old females: four independent samples from mifepristone-treated $S_132/UAS-dFOXO$ -TM adults (25 μ g mifepristone), and eight independent samples from controls (four of $S_132/UAS-dFOXO$ -TM without mifepristone and four of S_132 with mifepristone). Relative message abundance was measured by amplification with SYBR Green I on an ABI Prism 7700 cyclor (Applied Biosystems) and standardized against the *gap2dh* message. Differences between treatment and control were tested by nested analysis of variance, with replicate amplifications as a random effect.

Received 31 December 2003; accepted 6 April 2004; doi:10.1038/nature02549.

1. Clancy, D. J. *et al.* Extension of life-span by loss of CHICO, a *Drosophila* insulin receptor substrate protein. *Science* **292**, 104–106 (2001).
2. Tu, M.-P., Epstein, D. & Tatar, M. The demography of slow aging in male and female *Drosophila* mutant for the insulin-receptor substrate homolog *chico*. *Aging Cell* **1**, 75–80 (2002).
3. Tatar, M. *et al.* A mutant *Drosophila* insulin receptor homolog that extends life-span and impairs neuroendocrine function. *Science* **292**, 107–110 (2001).
4. Dillin, A., Crawford, D. K. & Kenyon, C. Timing requirements for insulin/IGF-1 signaling in *C. elegans*. *Science* **298**, 830–834 (2002).
5. Libina, N., Berman, J. R. & Kenyon, C. Tissue-specific activities of *C. elegans* DAF-16 in the regulation of lifespan. *Cell* **115**, 489–502 (2003).
6. Kenyon, C., Chang, J., Gensch, E., Rudner, A. & Tabtiang, R. A *C. elegans* mutant that lives twice as long as wild type. *Nature* **366**, 461–464 (1993).
7. Puig, O., Marr, M. T., Ruhf, M. L. & Tjian, R. Control of cell number by *Drosophila* FOXO: downstream and feedback regulation of the insulin receptor pathway. *Genes Dev.* **17**, 2006–2020 (2003).
8. Junger, M. A. *et al.* The *Drosophila* Forkhead transcription factor FOXO mediates the reduction in cell number associated with reduced insulin signaling. *J. Biol.* [online] **2**, Article 20; (<http://jbiol.com/content/2/3/20>) (2003).
9. Roman, G., Endo, K., Zong, L. & Davis, R. L. P[Switch], a system for spatial and temporal control of gene expression in *Drosophila melanogaster*. *Proc. Natl Acad. Sci. USA* **98**, 12602–12607 (2001).
10. Osterwalder, T., Yoon, K. S., White, B. H. & Keshishian, H. A conditional tissue-specific transgene expression system using inducible GAL4. *Proc. Natl Acad. Sci. USA* **98**, 12596–12601 (2001).
11. Dobens, L. L. *et al.* The *Drosophila* *bunched* gene is a homologue of the growth factor stimulated mammalian TSC-22 sequence and is required during oogenesis. *Mech. Dev.* **65**, 197–208 (1997).
12. Tatar, M., Bartke, A. & Antebi, A. The endocrine regulation of aging by insulin-like signals. *Science* **299**, 1346–1351 (2003).
13. Ashrafi, K. *et al.* Genome-wide RNAi analysis of *Caenorhabditis elegans* fat regulatory genes. *Nature* **421**, 268–272 (2003).
14. Garofalo, R. S. Genetic analysis of insulin signaling in *Drosophila*. *Trends Endocrinol. Metab.* **13**, 156–162 (2002).
15. Rulifson, E. J., Kim, S. K. & Nusse, R. Ablation of insulin-producing neurons in flies: growth and diabetic phenotypes. *Science* **296**, 1118–1120 (2002).
16. Cao, C. & Brown, M. R. Localization of an insulin-like peptide in brains of two flies. *Cell Tissue Res.* **304**, 317–321 (2001).
17. Brogiolo, W. *et al.* An evolutionarily conserved function of the *Drosophila* insulin receptor and insulin-like peptides in growth control. *Curr. Biol.* **11**, 213–221 (2001).
18. Wolkow, C. A., Kimura, K. D., Lee, M.-S. & Ruvkun, G. Regulation of *C. elegans* life-span by insulin-like signaling in the nervous system. *Science* **290**, 147–150 (2000).
19. Blüher, M., Kahn, B. B. & Kahn, R. C. Extended longevity in mice lacking the insulin receptor in adipose tissue. *Science* **299**, 572–574 (2003).
20. Blüher, M. *et al.* Adipose tissue selective insulin receptor knockout protects against obesity and obesity-related glucose intolerance. *Dev. Cell* **3**, 25–38 (2002).
21. Simon, A. F., Shih, C., Mack, A. & Benzer, S. Steroid control of longevity in *Drosophila melanogaster*. *Science* **299**, 1407–1410 (2003).
22. Gerisch, B., Weitzel, C., Kober-Eisermann, C., Rottiers, V. & Antebi, A. A hormonal signaling pathway influencing *C. elegans* metabolism, reproductive development, and life span. *Dev. Cell* **1**, 841–851 (2001).
23. Jia, K., Albert, P. S. & Riddle, D. L. DAF-9, a cytochrome P450 regulating *C. elegans* larval development and adult longevity. *Development* **129**, 221–231 (2002).
24. Murphy, C. T. *et al.* Genes that act downstream of DAF-16 to influence the lifespan of *Caenorhabditis elegans*. *Nature* **424**, 277–284 (2003).
25. Saucedo, L. J. *et al.* Rheb promotes cell growth as a component of the insulin/TOR signaling network. *Nature Cell Biol.* **5**, 566–571 (2003).
26. Vellai, T. *et al.* Influence of TOR kinase on lifespan in *C. elegans*. *Nature* **426**, 620 (2003).

Supplementary Information accompanies the paper on www.nature.com/nature.

Acknowledgements We thank F. Lemieux, A. Bente, M. Duon and M. Zerofsky for technical support, and A. Brunet, J. Wasserman, B. Edgar, G. Roman, O. Puig and T. Osterwalder for materials. This work was supported by the National Institute of Health (Institute of Aging), the American Federation of Aging Research, the Ellison Medical Foundation and a collaborative research grant from Pfizer.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to M.T. (Marc_Tatar@brown.edu).

The transcriptional programme of antibody class switching involves the repressor Bach2

Akihiko Muto¹, Satoshi Tashiro¹, Osamu Nakajima², Hideto Hoshino^{3*}, Satoru Takahashi³, Eiichirou Sakoda¹, Dai Ikebe¹, Masayuki Yamamoto^{3,4} & Kazuhiko Igarashi¹

¹Department of Biomedical Chemistry and Leukemia Program Project, Hiroshima University Graduate School of Biomedical Sciences, Hiroshima 734-8551, Japan

²Research Laboratory for Molecular Genetics, Yamagata University, Yamagata 990-9585, Japan

³Institute of Basic Medical Sciences,

⁴Center for Tsukuba Advanced Research Alliance and Exploratory Research for Advanced Technology, Environmental Response Project, University of Tsukuba, Tsukuba 305-8577, Japan

* Present address: National Institute of Advanced Industrial Science and Technology, Age Dimension Research Centre, Tsukuba 305-8566, Japan

Activated B cells differentiate to plasma cells to secrete IgM or, after undergoing class switch recombination (CSR), to secrete other classes of immunoglobulins^{1–4}. Diversification of antibody function by CSR is important for humoral immunity. However, it remains unclear how the decision for the bifurcation is made. Bach2 is a B-cell-specific transcription repressor interacting with the small Maf proteins whose expression is high only before the plasma cell stage^{5–7}. Here we show that Bach2 is critical for CSR and somatic hypermutation (SHM)^{2,4,8} of immunoglobulin genes. Genetic ablation of *Bach2* in mice revealed that Bach2 was required for both T-cell-independent and T-cell-dependent IgG responses and SHM. When stimulated *in vitro*, Bach2-deficient B cells produced IgM, as did wild-type cells, and abundantly expressed Blimp-1 (refs 9, 10) and XBP-1 (ref. 11), critical regulators of the plasmacytic differentiation¹², indicating that Bach2 was not required for the plasmacytic differentiation itself. However, they failed to undergo efficient CSR. These findings define Bach2 as a key regulator of antibody response and provide an insight into the orchestration of CSR and SHM during plasma cell differentiation.

After exposure to antigen *in vivo*, B-cell responses may involve SHM of the variable (V)-region exons to increase the affinity of antibody, and CSR of the constant (C)-region exons of immunoglobulin heavy chain (IgH) gene to produce antibodies with the same antigen specificity performing unique effector functions^{1–4,8}. Because SHM and CSR change DNA information and are therefore potentially mutagenic, their deployment must be tightly regulated. However, little is known about how transcription factors programme CSR and SHM in B cells.

Antigen-dependent terminal differentiation of B cells takes place within the secondary lymphoid organs such as spleen and lymph nodes^{2,4}. Using anti-Bach2 antibodies, we performed an immunohistochemical analysis of Bach2 in the mouse spleen. Bach2 was expressed in IgM-positive cells within the lymphoid follicles that were surrounded by marginal sinus expressing mucosal addressin cell adhesion molecule-1 (MadCAM-1 Fig. 1a). Marginal-zone B cells, which are IgM-positive and located outside MadCAM-1-expressing cells, were negative for Bach2. Bach2 was not detected in the CD3ε-positive T cells within the T-cell zone. Such an expression pattern suggests a role for Bach2 in antibody response.

To investigate this possibility, we disrupted the mouse *Bach2* gene by replacing the first coding exon with a *neo* resistance gene cassette (Supplementary Fig. 1). Breeding of *Bach2*^{+/-} mice resulted in *Bach2*^{-/-} progeny at the expected mendelian

frequency. *Bach2*^{-/-} progeny displayed no obvious abnormalities. No significant differences were observed for myeloid or T-cell numbers or cell-surface expression of markers including TER-119, Gr-1, Mac-1, CD4 and CD8 (data not shown). Numbers of pro-B (B220⁺, a pan-B-cell marker, and CD43⁺) and pre-B (B220⁺, CD43⁻, IgM⁻ and IgD⁻) cells in bone marrow were not affected in the *Bach2*^{-/-} mice (Fig. 1b, c). In contrast, the number of B220^{high} IgM⁺ recirculating B cells in the bone marrow of the *Bach2*^{-/-} mice was less than in 40% of control littermates. In spleen, naive B cells characterized as IgM⁺ IgD⁻ were present at normal levels in the *Bach2*^{-/-} mice. In contrast, numbers of mature B cells (B220⁺, IgM^{low} and IgD⁺) were significantly decreased in the *Bach2*^{-/-} mice (Fig. 1b, c). The absolute numbers of spleen cells in *Bach2*^{+/+} and *Bach2*^{-/-} mice were $(3.1 \pm 0.7) \times 10^8$ and $(2.1 \pm 0.4) \times 10^8$, respectively. Differentiation of both marginal zone B cells and follicular zone B cells was affected in the *Bach2*^{-/-} mice (Fig. 1d). In the spleens from *Bach2*-deficient mice, IgM⁺ B cells formed follicles that were surrounded by MadCAM-1-expressing marginal sinus (Fig. 1a, and data not shown). Thus, the formation of lymphoid follicles was obvious in the absence of *Bach2*. Staining with anti-*Bach2* antibodies was severely diminished in the spleens from *Bach2*-deficient mice, verifying the specificity of the *Bach2* signal in the wild-type mice (Fig. 1a). These results indicated that *Bach2* might be involved in the later stages of B-cell

development, in which antigen stimulation has a crucial role.

Serum immunoglobulin concentrations were measured by isotype-specific enzyme-linked immunosorbent assay (ELISA; Fig. 2a). The amounts of IgM in *Bach2*^{-/-} mice were fivefold those in wild-type mice. In contrast, the mean concentrations of IgG1, IgG2a, IgG2b, IgG3 and IgA were decreased to 31%, 5%, 33%, 4% and 25%, respectively, in *Bach2*^{-/-} mice relative to wild-type littermates. This altered pattern of serum immunoglobulin isotypes resembles that of hyper-IgM syndrome associated with immunodeficiency¹³. To examine whether *Bach2* is required for antigen-dependent humoral immune responses *in vivo*, the *Bach2*^{-/-} mice were immunized with T-cell-independent antigen 2,4-dinitrophenyl-conjugated Ficoll (DNP-Ficoll). The serum of *Bach2*^{-/-} mice before immunization contained higher levels of DNP-binding IgM antibody than that of the control mice (Fig. 2b). After immunization with DNP-Ficoll, DNP-binding IgM antibody increased in the wild-type mice but remained at similarly high levels in the *Bach2*^{-/-} mice. The *Bach2*^{+/+} mice mounted a specific IgG3 antibody response but the *Bach2*^{-/-} mice failed to do so (Fig. 2b). The T-cell-dependent antibody response was examined by immunizing 4-hydroxy-3-nitrophenylacetyl-conjugated chicken γ -globulin (NP-CGG). Whereas the *Bach2*^{+/+} mice showed a robust production of specific IgG1, the *Bach2*^{-/-} mice failed to do so (Fig. 2c). When immunized with sheep red blood cells (SRBC) (Fig. 2d), specific IgM antibodies

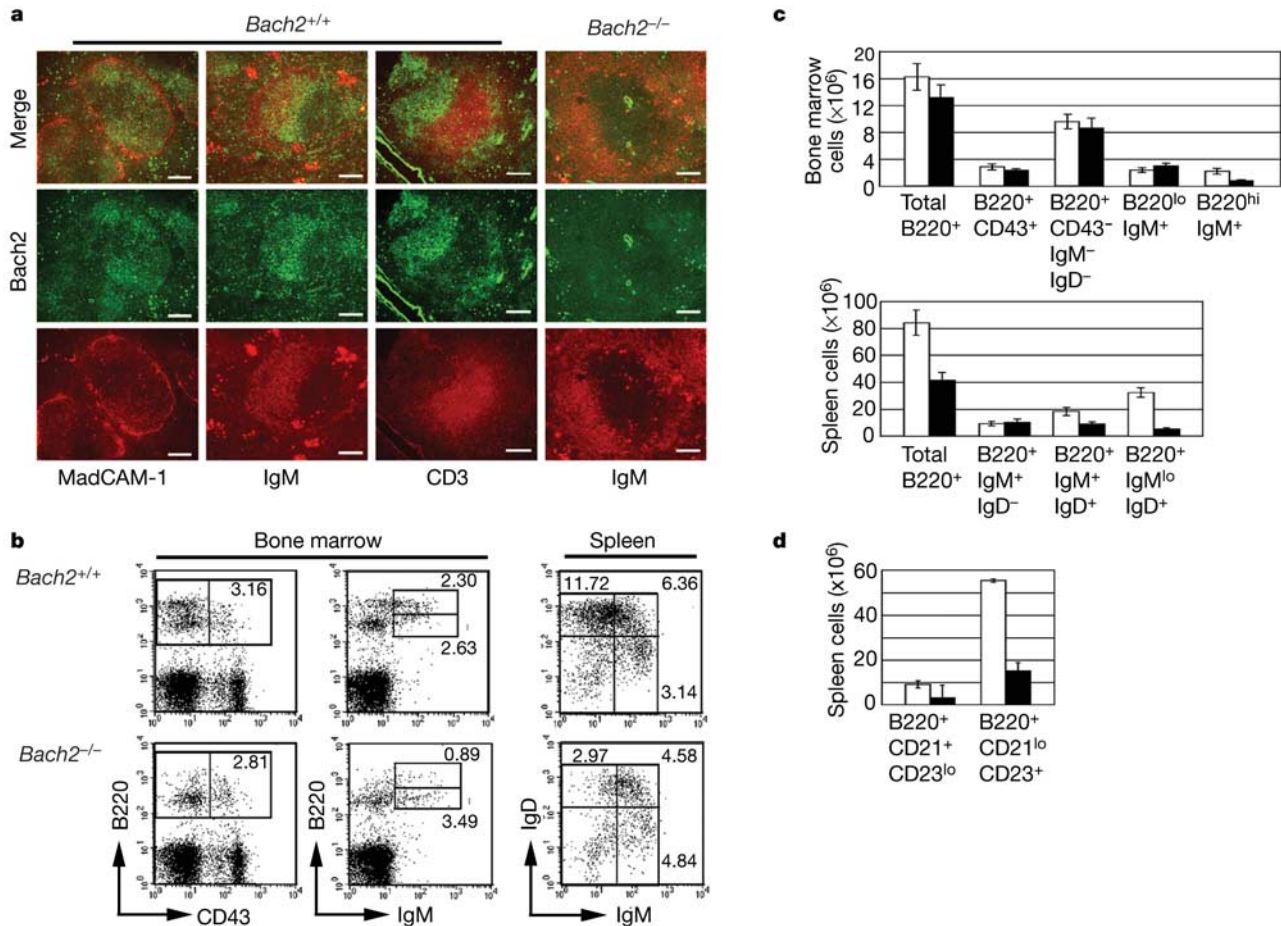


Figure 1 B-cell development in *Bach2*-deficient mice. **a**, Sections of spleen from *Bach2*^{+/+} (left) or *Bach2*^{-/-} (right) mice were double-stained for *Bach2* (middle row) and MadCAM-1, IgM or CD3e (lower row). Merged images (upper row) indicate *Bach2* (green) and MadCAM-1, IgM or CD3e (red). Scale bars, 100 μ m. **b**, Representative FACS analysis of bone marrow and spleen cells from *Bach2*^{+/+} and *Bach2*^{-/-} littermates. Percentages of gated cells are shown. **c**, Absolute cell numbers of B-cell

populations in bone marrow (upper) and in spleen (lower) from *Bach2*^{+/+} (open columns) and *Bach2*^{-/-} (filled columns) mice. Results are means $\pm$ s.e.m. for 8 bone marrow samples and 11 spleens from 4–11-week-old mice. **d**, Absolute cell numbers of marginal zone B (MZB: B220⁺, CD21⁺ and CD23^{lo}) and follicular zone B (FOB: B220⁺, CD21^{lo} and CD23⁺) cells from *Bach2*^{+/+} (open columns) and *Bach2*^{-/-} (filled columns) mice. Results are means $\pm$ s.e.m. for nine spleens from 6–11-week-old mice.

were robustly induced, indicating that antigen-driven differentiation of IgM plasma cells was intact in the *Bach2*^{-/-} mice. In contrast, specific IgG1 antibodies were not induced in the *Bach2*^{-/-} mice, again indicating defects in CSR. Thus, *Bach2* is essential for both T-cell-independent and T-cell-dependent production of isotype-specific antibodies. Because *Bach2*^{-/-} B cells produced cross-reacting low-affinity IgM antibodies without immunization (Fig. 2b, c), *Bach2* might be also required for normal checkpoint function of the B-cell receptor as well as for CSR.

Next we examined whether the *Bach2* deficiency affects the efficiency of SHM. Mice were immunized with NP-CGG, and complementary DNA was synthesized with RNAs extracted from mesenteric lymph node to determine the sequence of the VH186.2 exon, which is well characterized as the variable region encoding anti-NP hapten^{14,15}. Nucleotide exchange mutation frequency in wild-type and *Bach2*^{-/-} cells was 2.2×10^{-2} per base pair (bp) (18 clones were sequenced and 114 substitutions were found in 5,238 nucleotides) and 1.2×10^{-3} per bp (17 clones were sequenced and 6 substitutions were found in 4,947 nucleotides), respectively (Supplementary Fig. 2a, b). In *Bach2*^{-/-} mice, there were few nucleotide exchanges even in complementarity-determining regions that are critical for antigen binding (Fig. 2e). *Bach2*^{-/-} mice failed to select the known pattern of VDJ exons for high-affinity antibodies in secondary response (Supplementary Fig. 2c). These results demonstrated that *Bach2* is crucial for not only CSR but also for SHM.

Germinal centres (GCs) are microenvironments where antigen-

activated B cells undergo SHM and CSR with help from T cells². To investigate the involvement of *Bach2* in GC formation, spleens from *Bach2*^{-/-} and littermate mice were stained with peanut agglutinin (PNA) after immunization with SRBC. PNA⁺ GC B cells were clearly identified in the spleens from wild-type littermates. In contrast, GC B cells were not observed in the *Bach2*^{-/-} mice (Fig. 2f). The anatomical organization of IgM⁺ cells was comparable in *Bach2*^{-/-} and in control mice after immunization (Fig. 2f). These results indicated that *Bach2* is required for GC formation.

To determine directly whether *Bach2* deficiency affects CSR, we examined immunoglobulin production *in vitro* by treating spleen cells with lipopolysaccharide (LPS) and cytokines. The concentrations of IgG and IgA in the culture medium were significantly lower for the *Bach2*^{-/-} cells than for the wild-type cells (Fig. 3a). In contrast, *Bach2*^{-/-} spleen cells showed a relatively normal level of IgM secretion in comparison with the control cells, excluding the possibility that *Bach2* is required for antibody production in general. These results indicate that *Bach2* is dispensable for the development of IgM plasma cells but is crucial for the development of class-switched plasma cells.

Upon the initiation of CSR, combinations of the B-cell receptor and cytokine signals determine the specificity of CSR by inducing a particular class of IgH constant-region (C_H) gene transcript (germline IgH constant-region gene transcript, GLT)¹⁻⁴. After the completion of CSR, postswitch transcript (PST) containing intervening μ (I μ) and each of the switched C_H exons is expressed¹⁻⁴. To define

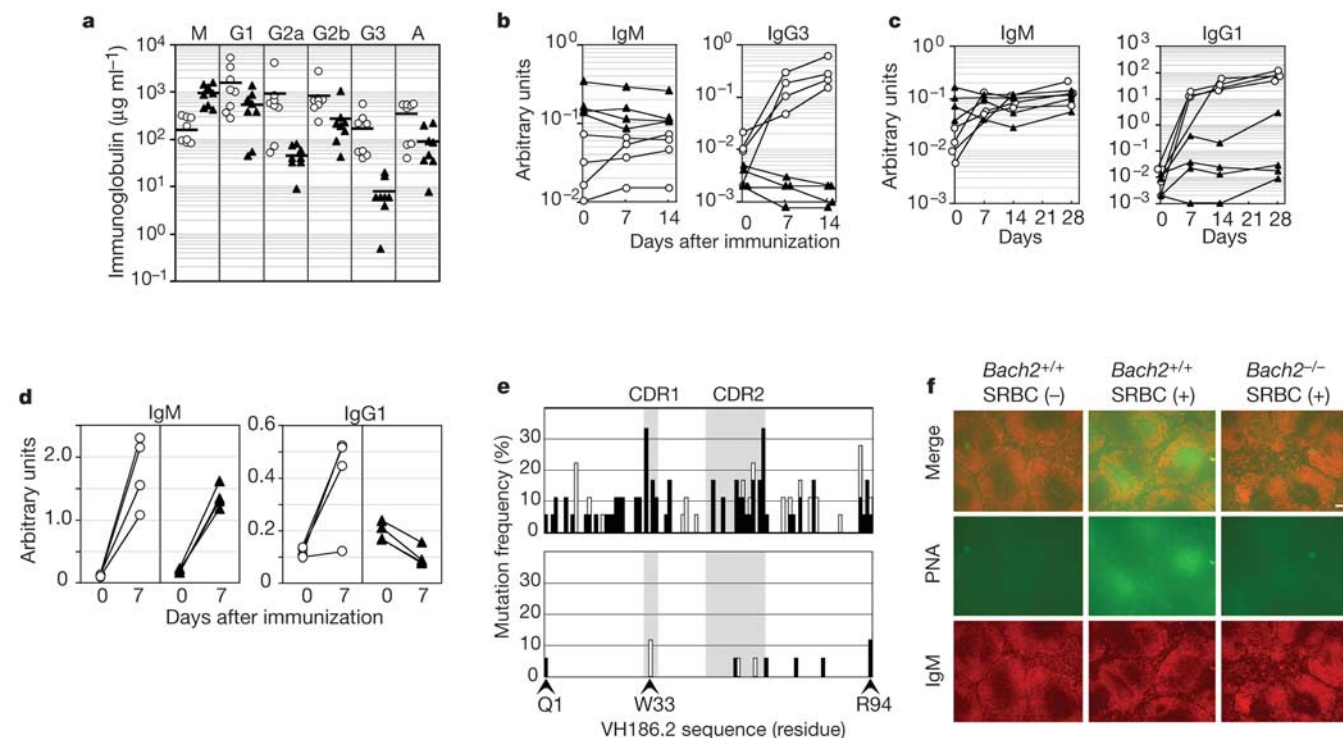


Figure 2 Impaired antibody production and SHM in *Bach2*-deficient mice. **a**, Immunoglobulin levels in sera from *Bach2*^{+/+} (open circles) and *Bach2*^{-/-} (filled triangles) 5–10-week-old mice. Horizontal bars indicate average concentrations. **b**, *Bach2*^{+/+} (open circles) or *Bach2*^{-/-} (filled triangles) mice were immunized with DNP-Ficoll. Anti-DNP IgM and IgG3 serum levels at 0, 7 and 14 days are shown in arbitrary units (A₄₁₀). **c**, *Bach2*^{+/+} (open circles) or *Bach2*^{-/-} (filled triangles) mice were immunized with NP-CGG on days 0 and 21. Anti-NP IgM and IgG1 serum levels at 0, 7, 14 and 28 days after immunization are shown. **d**, *Bach2*^{+/+} (open circles) or *Bach2*^{-/-} (filled triangles) mice were immunized with SRBC. Anti-SRBC IgM and IgG1 serum levels

at 0 and 7 days are shown. **e**, Mutation frequencies in the VH186.2 transcripts of the μ isotype from mesenteric lymph node of NP-CGG-immunized *Bach2*^{+/+} (upper panel) and *Bach2*^{-/-} (lower panel) mice. Replacements (filled bars) or silent mutations (open bars) at specific positions are plotted against residue number from 1 to 94. CDR, complementarity-determining region. **f**, Sections of mouse spleen from unimmunized *Bach2*^{+/+} (left), SRBC-immunized *Bach2*^{+/+} (centre) and SRBC-immunized *Bach2*^{-/-} (right) mice were double-stained for PNA (middle row) and IgM (bottom row). Merged images are shown in the top row. Scale bar, 100 μ m.

the step in CSR that requires Bach2, we examined GLT and PST in B cells stimulated with LPS and cytokines by RT-PCR. After 2 days in culture, the levels of $\gamma 2b$, $\gamma 3$ and α GLTs in the *Bach2*^{-/-} B cells were comparable to those in the wild-type B cells (Fig. 3b). In contrast, the quantity of $\gamma 1$ GLT was decreased in the *Bach2*^{-/-} B cells. Thus, the activation of chromatin structure in the switch regions is largely unaffected by the absence of Bach2 except for $\gamma 1$ GLT. In contrast to GLT, PSTs were severely decreased in *Bach2*^{-/-} B cells (Fig. 3c). These results indicate that Bach2 deficiency caused a common defect in the cleavage and/or ligation of the switch regions. Because cell proliferation is a prerequisite for CSR¹⁶, proliferation activities of *Bach2*^{-/-} B cells were assessed by measuring BrdU incorporation

by B220⁺ B cells that were purified from spleen. Responses to LPS, which was used to induce plasma cells in the above experiments, were indistinguishable between *Bach2*^{-/-} and wild-type spleen B cells (Supplementary Fig. 3).

Bach2 might regulate the expression of genes involved in CSR and SHM. This possibility was investigated by comparing the expression of activation-induced cytidine deaminase (AID), which is necessary for both SHM and CSR¹⁷. Using purified B220⁺ spleen B cells stimulated with LPS, quantitative polymerase chain reaction with reverse transcription (RT-PCR) revealed that the expression of AID was reduced to 0.12% in the *Bach2*^{-/-} B cells after 2 days of treatment with LPS (Fig. 3d). Considering its well-characterized crucial role in SHM and CSR, we next investigated whether the reduced expression of AID caused the CSR defects of *Bach2*^{-/-} B cells by retroviral complementation. *Bach2*^{-/-} spleen B cells treated with LPS and interleukin-4 were infected with retroviruses expressing both AID and enhanced green fluorescent protein (EGFP) or EGFP alone as a control. The fraction of isotype-switched cells among the EGFP-positive AID-expressing cells was determined by the expression of surface IgG1 and IgG3 on the third day after infection. As shown in Fig. 4a, we did not observe robust restoration of CSR in *Bach2*^{-/-} B cells by AID delivery. These results suggest that Bach2-deficient B cells suffer from not only the reduced expression of AID but also de-regulation of other genes involved directly or indirectly in CSR.

This possibility was investigated by comparing the expression of genes implicated in plasma cell development. As shown in Fig. 3d, we observed higher expression levels of transcription factors that drive the terminal differentiation of B cells to plasma cells (Blimp-1 (refs 9, 10), XBP-1 (ref. 11) and IRF4 (ref. 18)). We observed almost eightfold higher Blimp-1 expression in the *Bach2*^{-/-} B cells than in *Bach2*^{+/+} cells on day 2. The expression of BCL-6 (ref. 19) and OCA-B^{20,21}, which are required for GC formation, was respectively largely unaffected and somewhat increased in the absence of Bach2. Pax5 (ref. 22), c-Myc²³ and c-Fos²⁴ are normally downregulated after terminal differentiation of B cells. The expression of Pax5 showed a more rapid decline in the *Bach2*^{-/-} B cells. The overexpression of Blimp-1 might be a cause of the decrease in Pax5 and AID expression because Blimp-1 drives plasmacytic differentiation by suppressing mature B-cell-associated genes, including Pax5 and AID^{9,12}. Thus, the pattern of gene expression in *Bach2*^{-/-} B cells stimulated *in vitro* was coherent with plasma cell differentiation.

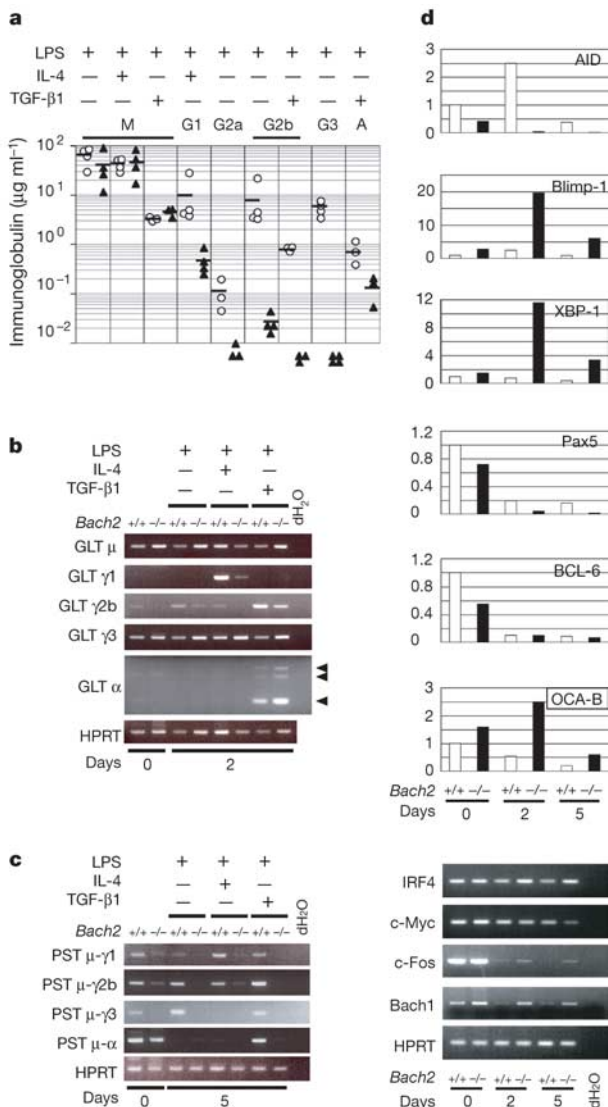


Figure 3 Impaired CSR in the *Bach2*-deficient B cells. **a**, RBC-depleted splenocytes from *Bach2*^{+/+} (open circles) and *Bach2*^{-/-} (filled triangles) mice were cultured with LPS and the indicated cytokines for 7 days, and secreted immunoglobulin levels were measured. **b**, Expression of GLT containing I and C_H exons of each isotype in splenocytes stimulated with LPS and indicated cytokines for 2 days. GLT α is indicated by arrowheads. dH₂O, distilled water. **c**, Expression of PST containing I μ and each C_H exon in splenocytes stimulated with LPS and the indicated cytokines for 5 days. **d**, Expression of B-cell-maturation-related genes was determined with purified spleen B cells stimulated with LPS for 0, 2 or 5 days. The expression profiles were analysed by quantitative real-time RT-PCR, with HPRT expression as the normalization control. Averages of two independent cultures from *Bach2*^{+/+} (open columns) and *Bach2*^{-/-} (filled columns) mice are shown.

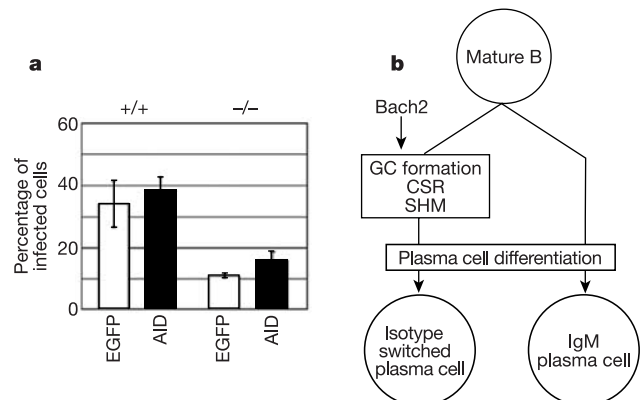


Figure 4 Characterization of genetic circuitry involving Bach2. **a**, Flow cytometric analysis of surface IgG1 and IgG3 expression in *Bach2*^{+/+} (left) and *Bach2*^{-/-} (right) B cells infected with control (open columns) or AID (filled columns) viruses. Results are means $\pm$ s.e.m. for three independent experiments. **b**, Model for Bach2 function in activated B cells. Bach2 has a crucial function in progression through GC formation, CSR and SHM before terminal differentiation into plasma cells. Generation of IgM plasma cells is independent of Bach2. Plasmacytic differentiation depends on Blimp-1 and XBP-1.

Bach1 (refs 6, 25, 26) might partly replace Bach2 because it is expressed in B cells (Fig. 3d).

The differentiation of plasma cells poses a general question about lineage choice in haematopoietic cells in that CSR, which is not reversible, is selected only in a subset of plasma cells. Studies on lineage choice have indicated that transcription factors might programme individual lineages. We have shown here that antibody responses involving CSR, SHM and GC formation were severely impaired in Bach2-deficient mice (Fig. 4b). In addition, *Bach2*^{-/-} B cells can differentiate to IgM-secreting plasma cells. In this regard, Bach2 stands in contrast with known plasma-cell transcription factors such as Blimp-1 and XBP-1, which are necessary for both non-IgM and IgM plasma cells^{10,11}. Our observations demonstrate that Bach2 is a hitherto unknown regulator of antibody response. Understanding the direct target genes of Bach2 as well as the signalling cascades that regulate the activity of Bach2 in B cells could reveal a novel molecular link between B-cell activation and antibody response. □

Methods

Immunofluorescence staining

Frozen sections of spleen were stained with antibodies: FITC-conjugated anti-mouse IgM, biotin-conjugated anti-mouse IgM, biotin-conjugated anti-mouse CD3ε (145-2C11) and purified anti-mouse MadCAM-1 (MECA-367) (BD Pharmingen), and with FITC-conjugated PNA (HONEN). Cy3-conjugated goat anti-rat IgG (Rockland), Alexa Fluor 488 goat anti-rabbit IgG (Molecular Probes) and Cy3-conjugated avidin were used as secondary antibodies. Image acquisition was as described previously²⁷.

Disruption of the *Bach2* gene

Analysis of genomic phage clones of mouse *Bach2* was as described previously²⁸; detailed information is available from the authors on request. The *Bach2*-targeting vector (pB2TV) was constructed by using DNA fragments containing the β-galactosidase (*lacZ*) gene with SV40 polyadenylation (poly(A)) signal derived from pCMVβ vector (Clontech) and the neomycin resistance (*neo*) gene driven by the phosphoglycerate kinase (PGK)-1 promoter (PGK-*neo*), which was flanked by a loxP sequence on both sides. These cassettes were cloned between short-arm and long-arm DNA. The 3.0-kb short-arm DNA was isolated from the genomic phage clone by PCR with a *Bach2*-specific primer and a primer for the vector sequence. The *Bach2*-specific primer possessed an *EcoRI* site at the 5' end (B2SAR primer 5'-AATTAgaattcGTTCAACCTGGAAATAAGAG-3') and the PCR product was cloned as an *EcoRI* DNA. The 6.7-kb long-arm DNA covered between the *EcoRI* site and the 3' end of the phage clone end. A cDNA containing diphtheria toxin A driven by the PGK promoter (PGK-DTA), without a poly(A) signal, was then subcloned beyond the short arm to complete the construct (see Supplementary Fig. 1a). The pB2TV was linearized with *NorI*. The linearized targeting vector was transfected into embryonic stem cell line E14. Appropriately targeted embryonic stem cell clones (Supplementary Fig. 1b) were used to obtain chimaeric mice that transmitted the disrupted locus through the germ line. Two independent embryonic stem clones gave rise to *Bach2*^{-/-} mouse lines with the same phenotype. The 4–12-week-old mice used in this study were backcrossed to the C57BL/6J background for one or two generations; wild-type littermates were used as controls. Immunoblotting of bone marrow extracts with anti-Bach2 antibody detected Bach2 in the wild-type mice but not in *Bach2*^{-/-} mice (Supplementary Fig. 1c).

Fluorescence-activated cell sorting (FACS) analysis

FACS analysis was performed with antibodies against mouse CD45R/B220, CD43 (S7), IgM (R6-60.2), IgD, IgG1, IgG3, TER-119, Ly-6G/Gr-1, CD11b/Mac-1, CD4, CD8, CD21 and CD23 (BD Pharmingen). The cells were analysed on a FACScalibur with CellQuest software (Becton Dickinson).

Determination of immunoglobulin levels

Immunoglobulin concentrations were determined with a standard procedure by using antibodies specific for each mouse immunoglobulin isotype and with *p*-nitrophenylphosphate as a substrate for the alkaline phosphatase-conjugated secondary antibodies (Clonotyping System/AP; Southern Biotechnology Associates).

In vivo immunizations and SHM assay

Mice (7–12 weeks old, four mice for each genotype) were immunized intraperitoneally with 100 μg of DNP-Ficoll (Biosearch Technologies) in PBS (Nissui), 100 μg NP-chicken gamma-globulin (NP-CGG; Biosearch Technologies) mixed with alum, or 10⁶ cells of SRBC. Levels of anti-DNP, anti-NP and anti-SRBC antibodies were determined by ELISA with DNP-BSA, NP-BSA (Biosearch Technologies) and soluble SRBC, respectively, as capture agents. Somatic hypermutation assay was performed as described previously¹⁷.

In vitro class switching assay

RBC-depleted splenocytes (10⁵) were cultured in 96-well plates in a total volume of 200 μl RPMI medium. The cells were treated with 20 μg ml⁻¹ LPS (0111:B4; Sigma), 10 ng ml⁻¹ recombinant mouse interleukin-4 (BD Pharmingen) and 1 ng ml⁻¹ recombinant human TGF-β1 (R&D Systems) in indicated combinations. Proliferation of spleen B cells (B220⁺

magnetic bead selection; Miltenyi Biotech) were measured by using cell-proliferation ELISA (BrdU kit; Roche).

RT-PCR

RNA preparation and cDNA synthesis were as described previously⁷. Sequences of PCR primers are available from the authors on request. For quantitative TaqMan RT-PCR (AID, Blimp-1, BCL-6, Pax5, XBP-1 and hypoxanthine guanine phosphoribosyl transferase (HPRT)), cDNA samples were analysed in triplicate with a TaqMan universal PCR master kit and an ABI 7700 sequence detection system with primers and probes from Assays-on-Demand gene expression products (Applied Biosystems). For quantitative RT-PCR (OCA-B and HPRT), cDNA samples were analysed with a LightCycler FastStart DNA master SYBR Green I kit and a LightCycler system (Roche).

Retroviral infection

Preparation of retrovirus was as described previously²⁷. Purified B220⁺ spleen B cells were preactivated with 20 μg ml⁻¹ LPS and 10 ng ml⁻¹ interleukin-4 for 2 days before infection. Preactivated cells were suspended at a density of 5 × 10⁵ cells ml⁻¹ in the medium containing retrovirus supplemented with 16 μg ml⁻¹ Polybrene, centrifuged at 927g for 90 min at 32 °C and incubated for 48 h. The cells were then washed and incubated for 24 h. To analyse the expression of surface IgG1 and IgG3 by FACS, live cells were gated on the basis of cell size and staining by propidium iodide.

Received 10 February; accepted 21 April 2004; doi:10.1038/nature02596.

Published online 19 May 2004.

1. Snapper, C., Marcu, K. & Zelazowski, P. The immunoglobulin class switch: beyond 'accessibility'. *Immunity* **6**, 217–223 (1997).
2. Kinoshita, K. & Honjo, T. Linking class-switch recombination with somatic hypermutation. *Nature Rev. Mol. Cell Biol.* **2**, 493–503 (2001).
3. Manis, J., Tian, M. & Alt, F. Mechanism and control of class-switch recombination. *Trends Immunol.* **23**, 31–39 (2002).
4. Honjo, T., Kinoshita, K. & Muramatsu, M. Molecular mechanism of class switch recombination: linkage with somatic hypermutation. *Annu. Rev. Immunol.* **20**, 165–196 (2002).
5. Igarashi, K. *et al.* Regulation of transcription by dimerization of erythroid factor NF-E2p45 with small Maf proteins. *Nature* **367**, 568–572 (1994).
6. Oyake, T. *et al.* Bach proteins belong to a novel family of BTB-basic leucine zipper transcription factors that interact with MafK and regulate transcription through the NF-E2 site. *Mol. Cell. Biol.* **16**, 6083–6095 (1996).
7. Muto, A. *et al.* Identification of Bach2 as a B-cell-specific partner for small maf proteins that negatively regulate the immunoglobulin heavy chain gene 3' enhancer. *EMBO J.* **17**, 5734–5743 (1998).
8. Papavasiliou, F. & Schatz, D. Somatic hypermutation of immunoglobulin genes: merging mechanisms for genetic diversity. *Cell* **109**, S35–S44 (2002).
9. Shaffer, A. *et al.* Blimp-1 orchestrates plasma cell differentiation by extinguishing the mature B cell gene expression program. *Immunity* **17**, 51–62 (2002).
10. Shapiro-Shelef, M. *et al.* Blimp-1 is required for the formation of immunoglobulin secreting plasma cells and pre-plasma memory B cells. *Immunity* **19**, 607–620 (2003).
11. Reimold, A. *et al.* Plasma cell differentiation requires the transcription factor XBP-1. *Nature* **412**, 300–307 (2001).
12. Calame, K., Lin, K. & Tunyaplin, C. Regulatory mechanisms that determine the development and function of plasma cells. *Annu. Rev. Immunol.* **21**, 205–230 (2003).
13. Durandy, A. & Honjo, T. Human genetic defects in class-switch recombination (hyper-IgM syndromes). *Curr. Opin. Immunol.* **13**, 543–548 (2001).
14. Allen, D., Simon, T., Sablitzky, F., Rajewsky, K. & Cumano, A. Antibody engineering for the analysis of affinity maturation of an anti-hapten response. *EMBO J.* **7**, 1995–2001 (1989).
15. Jacob, J., Kelsø, G., Rajewsky, K. & Weiss, U. Intracloonal generation of antibody mutants in germinal centres. *Nature* **354**, 389–392 (1991).
16. Hasbold, J., Lyons, A., Kehry, M. & Hodgkin, P. Cell division number regulates IgG1 and IgE switching of B cells following stimulation by CD40 ligand and IL-4. *Eur. J. Immunol.* **28**, 1040–1051 (1998).
17. Muramatsu, M. *et al.* Class switch recombination and hypermutation require activation-induced cytidine deaminase (AID), a potential RNA editing enzyme. *Cell* **102**, 553–563 (2000).
18. Mittrucker, H. *et al.* Requirement for the transcription factor LSIRF/IRF4 for mature B and T lymphocyte function. *Science* **275**, 540–543 (1997).
19. Ye, B. *et al.* The BCL-6 proto-oncogene controls germinal-centre formation and Th2-type inflammation. *Nature Genet.* **16**, 161–170 (1997).
20. Schubart, D., Rolink, A., Kosco-Vilbois, M., Botteri, F. & Matthias, P. B-cell-specific coactivator OBF-1/OCA-B/Bob1 required for immune response and germinal centre formation. *Nature* **383**, 538–542 (1996).
21. Kim, U. *et al.* The B-cell-specific transcription coactivator OCA-B/OBF-1/Bob-1 is essential for normal production of immunoglobulin isotypes. *Nature* **383**, 542–547 (1996).
22. Barberis, A., Widenhorn, K., Vitelli, L. & Busslinger, M. A novel B-cell lineage-specific transcription factor present at early but not late stages of differentiation. *Genes Dev.* **4**, 849–859 (1990).
23. Lin, Y., Wong, K. & Calame, K. Repression of c-myc transcription by Blimp-1, an inducer of terminal B cell differentiation. *Science* **276**, 596–599 (1997).
24. Monroe, J. Up-regulation of c-fos expression is a component of the mIg signal transduction mechanism but is not indicative of competence for proliferation. *J. Immunol.* **140**, 1454–1460 (1988).
25. Sun, J. *et al.* Heme regulates the dynamic exchange of Bach1 and NF-E2-related factors in the Maf transcription factor network. *Proc. Natl Acad. Sci. USA* **101**, 1461–1466 (2004).
26. Brand, M. *et al.* Dynamic changes in transcription factor complexes during erythroid differentiation revealed by quantitative proteomics. *Nature Struct. Mol. Biol.* **11**, 73–80 (2004).
27. Muto, A. *et al.* Activation of Maf/AP-1 repressor Bach2 by oxidative stress promotes apoptosis and its interaction with promyelocytic leukemia nuclear bodies. *J. Biol. Chem.* **277**, 20724–20733 (2002).
28. Sun, J. *et al.* The promoter of mouse transcription repressor *bach1* is regulated by Sp1 and trans-activated by Bach1. *J. Biochem. (Tokyo)* **130**, 385–392 (2001).

Supplementary Information accompanies the paper on www.nature.com/nature.

Acknowledgements We thank M. Kobayashi, D. Engel, M. Muramatsu and T. Honjo for discussion and comments on the manuscript, M. Busslinger for information on RT-PCR analysis, N. Kaneko for the preparation of tissue sections, and M. Kanno for the use of FACS machinery. This work was supported by Grants-in-aid from the Ministry of Education, Culture, Sport, Science and Technology of Japan and Hiroshima University 21st century COE programme. A.M. was initially supported by the Research Fellowships for Young Scientists from the Japanese Society for Promotion of Science.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to K.I. (igarak@hiroshima-u.ac.jp).

Intergenic transcription is required to repress the *Saccharomyces cerevisiae* *SER3* gene

Joseph A. Martens, Lisa Laprade & Fred Winston

Department of Genetics, Harvard Medical School, 77 Avenue Louis Pasteur, Boston, Massachusetts 02115, USA

Transcription by RNA polymerase II in *Saccharomyces cerevisiae* and in humans is widespread, even in genomic regions that do not encode proteins^{1–6}. The purpose of such intergenic transcription is largely unknown, although it can be regulatory^{7,8}. We have discovered a role for one case of intergenic transcription by studying the *S. cerevisiae* *SER3* gene. Our previous results demonstrated that transcription of *SER3* is tightly repressed during growth in rich medium⁹. We now show that the regulatory region of this gene is highly transcribed under these conditions and produces a non-protein-coding RNA (*SRG1*). Expression of the *SRG1* RNA is required for repression of *SER3*. Additional experiments have demonstrated that repression occurs by a transcription-interference mechanism in which *SRG1* transcription across the *SER3* promoter interferes with the binding of activators. This work identifies a previously unknown class of transcriptional regulatory genes.

The *S. cerevisiae* *SER3* gene encodes a phosphoglycerate dehydrogenase that catalyses a step in serine biosynthesis¹⁰. To investigate *SER3* repression in greater detail, we performed chromatin immunoprecipitation (ChIP) experiments to test for the association of general transcription factors with the *SER3* regulatory region *in vivo*. Surprisingly, our results showed that under repressing conditions, a significant level of both TATA-binding protein (TBP) and RNA polymerase II (Pol II) bind 5' of the *SER3* TATA element (Fig. 1a). Additional ChIP experiments showed that other factors, including phosphorylated forms of Pol II and mRNA capping factors, are also associated with this region (unpublished data). Thus, when *SER3* is repressed, its regulatory region is associated with factors required for active transcription. As described below, this region contains a second TATA element (shown in Fig. 1a).

To determine whether transcription occurs in the *SER3* regulatory region, we performed transcription run-on experiments. These results showed that there is a high level of active transcription (Fig. 1b, probes 3–5). This transcription occurs on the same strand as *SER3* as determined by its hybridization specificity for single-stranded probes.

We gained additional insight into the transcription occurring 5' of *SER3* by comparing the DNA sequence of this region to that of four yeasts closely related to *S. cerevisiae*^{11,12}. This analysis

(Supplementary Fig. 1) revealed that, in addition to the TATA element proximal to the *SER3* coding region (–103 relative to the *SER3* ATG), there is a second TATA element at –558 that is perfectly conserved in all five yeasts. Short conserved sequences 5' of this TATA element suggest the presence of regulatory sites. Within the transcribed region, there are three conserved sequence motifs between –262 and –156. Deletion analysis strongly suggests that this region functions as a *SER3* upstream activating sequence (UAS) (Supplementary Fig. 2). Outside of these sequence elements there is no significant conservation or open reading frame in this region, suggesting that it does not encode a protein. Because this region is transcribed and, as will be shown, has an identified function, we have designated it as the *SRG1* gene (*SER3* regulatory gene 1). We will refer to the TATA element at –558 as the *SRG1* TATA element.

To test for a role for *SRG1* in *SER3* repression, we constructed a mutation in the *SRG1* TATA sequence, from TATAAA to CCTAGG (*srG1-1*). By northern analysis, the *SRG1* transcript in a wild-type strain appears primarily as a diffuse band of approximately 550 bases (Fig. 2a). A longer RNA transcript is also present at a lower level whose length is consistent with initiation at the *SRG1* initiation site and readthrough across *SER3*. In the *srG1-1* TATA mutant, both RNAs are undetectable. Significantly, *SER3* transcription is derepressed to a high level in this mutant, demonstrating an

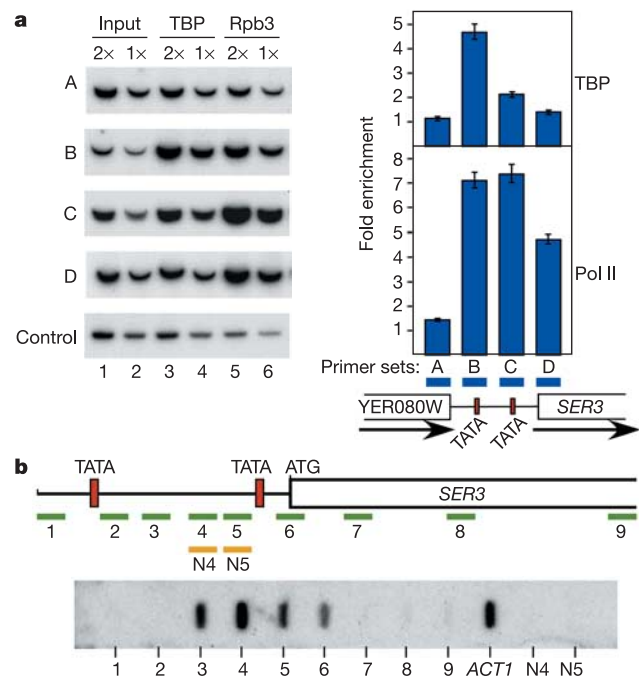


Figure 1 Evidence for active transcription 5' of *SER3*. **a**, ChIP analysis of TBP and Pol II (Rpb3) was performed on wild-type strain FY2245 grown in repressing conditions (YPD medium). A representative set of PCR reactions from two-fold dilutions of chromatin is shown. The regions amplified by PCR (A–D) are marked by blue bars in the diagram of the *SER3* promoter on the right. The control primer set amplifies a region of chromosome V that lacks open reading frames²⁵. The graphs summarize the results of three independent experiments, with each value representing the average and standard error of the fold enrichment (ratio of the percentage of *SER3* immunoprecipitated to the percentage of the control region immunoprecipitated). **b**, Transcription run-on analysis at *SER3*. Transcription run-on analysis was performed on wild-type strain FY2097 grown in repressing conditions. A schematic of *SER3* is shown, with the green bars representing antisense oligonucleotides that detect transcription from the Watson strand. The two yellow bars (N4, N5) represent sense oligonucleotides used as negative controls. An oligonucleotide that detects transcription of *ACT1* was included as a positive control.

essential role for *SRG1* in the repression of *SER3*.

Characterization of the *SRG1* transcript was achieved through several experiments. First, primer-extension analysis showed that the *SRG1* RNA has discrete 5' ends positioned 78 and 79 bases 3' of the *SRG1* TATA (Fig. 2b). Second, to test whether the *SRG1* RNA is polyadenylated, we compared the levels of poly(A) enriched and total RNA for *SRG1*, *ACT1* (a polyadenylated control) and *SNR190* (a non-polyadenylated control¹³). Our results (Supplementary Fig. 3) show that *SRG1* is significantly enriched in the poly(A) sample, which strongly suggests that it is polyadenylated. Third, to test whether the presence of TBP and Pol II in the *SER3* 5' region (Fig. 1a) is due to their association with the actively transcribed *SRG1* gene, we performed ChIP experiments, comparing wild-type and *srg1-1* strains. These results show that the level of both TBP and Pol II are significantly decreased over the *SRG1* TATA region (region B) in the *srg1-1* mutant (Fig. 2c). The levels are not reduced over regions C and D, owing to the derepression of *SER3* in the *srg1-1* mutant. Taken together, our results identify *SRG1* as a Pol-II-dependent, highly expressed, non-coding RNA (Fig. 2d).

We considered three models for the role of *SRG1* in *SER3* repression: an RNA-mediated model, a promoter-competition model and a transcription-interference model. In the RNA-mediated model, the *SRG1* RNA product itself is required for repression of *SER3*, perhaps by interacting with proteins.

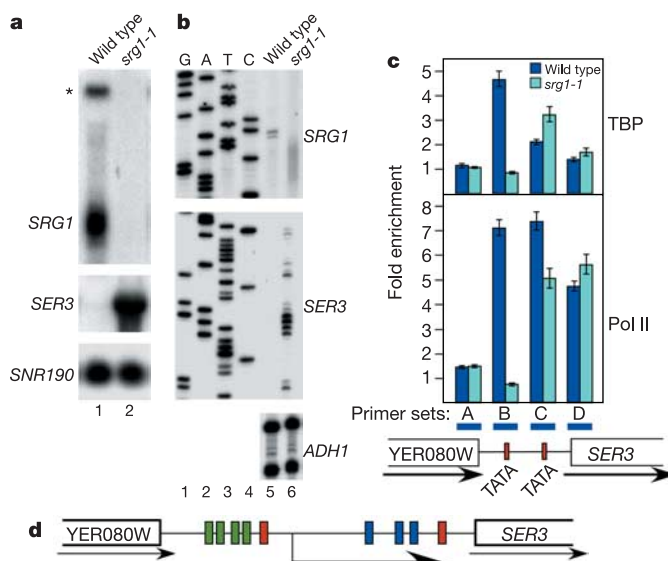


Figure 2 Characterization of *SRG1* transcription. **a**, Northern analysis of *SRG1* and *SER3* was performed on wild-type (FY2097) and *srg1-1* (FY2250) strains. *SNR190* was a loading control. The asterisk marks a minor RNA product of approximately 2 kb that is consistent with a readthrough product from the *SRG1* promoter to the end of *SER3*. This RNA is also detected with a *SER3* probe after long exposure (data not shown). **b**, Primer-extension analysis of *SRG1* and *SER3* was performed on wild-type (FY2256) and *srg1-1* (FY2247) strains. The two transcription start sites for *SRG1* (lane 5; top panel) map to 78 and 79 bases 3' of the *SRG1* TATA. *SER3* transcription initiates at multiple sites within a 28 base region that begins 52 bases 3' of the *SER3* TATA (lane 6, middle panel). The level of *ADH1* mRNA was measured as a control for the amount of RNA used in each reaction (bottom panel). **c**, ChIP analysis for TBP and Pol II (Rpb3) performed on wild-type (FY2245) and *srg1-1* (FY2247) strains. The values shown are calculated from three independent experiments as described in the legend of Fig. 1a. **d**, A diagram of *SRG1*. The dark arrow represents the approximate position of the *SRG1* transcript. The coloured boxes represent conserved sequences in this region. The red boxes are TATA elements, the green boxes are putative UAS elements for *SRG1*, and the blue boxes are putative UAS elements for *SER3* (see Supplementary Fig. 2).

There are several precedents for RNA-mediated control of transcription, including X chromosome inactivation in mammals and dosage compensation in *Drosophila*¹⁴. In the promoter-competition model, the *SRG1* promoter outcompetes that of *SER3* for transcription factors, inhibiting activation of *SER3*. Evidence for this mechanism has been previously described^{15,16}. In the transcription-interference model, transcription of *SRG1* across the *SER3* promoter region blocks the binding of transcription factors necessary for *SER3* activation. Transcription interference has previously been shown to occur either as a regulatory mechanism or as a consequence of deletion of the transcription terminator between genes^{17–23}.

First, if the *SRG1* RNA product functions in repression, it may act in *trans*. Therefore, we tested the ability of *SRG1* to repress when it is located either in *cis* or in *trans* to *SER3*. We constructed diploid strains in which one copy of the *SER3*-coding sequence was replaced by the sequence for orotidine-5'-phosphate decarboxylase (*URA3*) (*ser3Δ::URA3*), allowing expression from each *SER3* promoter to be assayed. One diploid had the *srg1-1* mutation adjacent to wild-type *SER3* (Fig. 3a, FY2254) and the other had *srg1-1* adjacent to *ser3Δ::URA3* (Fig. 3a, FY2255). Northern analysis of these strains demonstrates that wild-type *SRG1* represses the *SER3* promoter only when it is in *cis* (Fig. 3a). These results rule out the possibility that the *SRG1* RNA product represses transcription as a *trans*-acting factor, although they do not rule out an RNA-mediated model in which the RNA product acts in *cis*.

Next, we tested the promoter-competition model using two derivatives of *SRG1*. In each construct, we replaced 150 base pairs

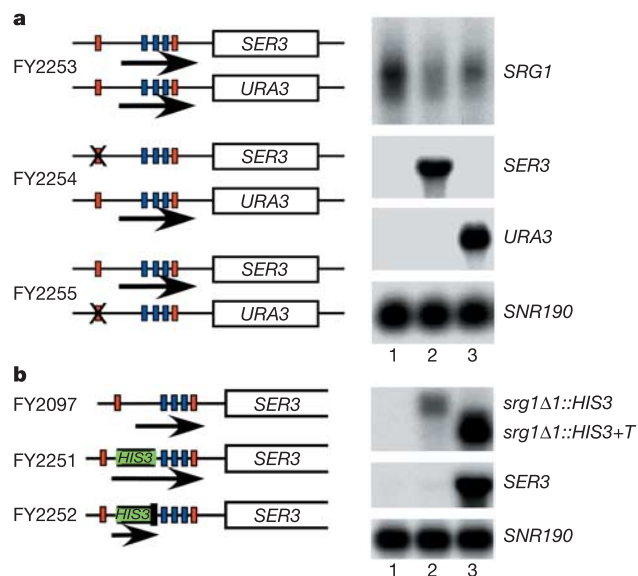


Figure 3 Tests of models for *SRG1* function. **a**, A *cis/trans* test for *SRG1* function. The left part of the figure shows a diagram of the *SER3* alleles of the three diploids used in the experiment (FY2253, FY2254 and FY2255). The alleles that contain the *srg1-1* TATA mutation are indicated by an X over the *SRG1* TATA box. The black arrows represent *SRG1* transcripts. The right part is a northern analysis of each diploid for *SRG1*, *SER3* and *URA3* RNA levels, with *SNR190* as a loading control. **b**, A test of the promoter-competition model. The left part shows a diagram of the *SRG1* allele (FY2097) and two different *HIS3* insertions (the green boxes): one without (FY2251) and one with its terminator sequence (black box; FY2252). The red boxes represent the *SRG1* and *SER3* TATAs. The blue boxes represent three sequence motifs conserved between multiple yeast strains that may serve as a *SER3* UAS (see Supplementary Fig. 1). The black arrows represent transcripts initiated from the *SRG1* promoter. On the right is a northern analysis of *SRG1*, *SER3* and *SNR190* (loading control) from the strains shown on the left. The transcripts generated from the *SRG1* promoter were detected using a probe for *HIS3* (lanes 2 and 3; top panel).

(bp) of *SRG1* sequence with the *HIS3* reporter gene coding sequence such that the putative *SER3* UAS remained intact (Fig. 3b). In one construct (FY2251), the *HIS3* transcription-termination region was not included, allowing *SRG1* transcription to elongate across the *SER3* UAS as in a wild-type strain. In the other construct (FY2252), the *HIS3* transcription-termination sequence was included, thereby terminating *SRG1* transcription 5' of the *SER3* promoter. Northern analysis confirmed that *SRG1* transcription behaved as expected in each *HIS3* insertion (Fig. 3b). When *SER3* mRNA levels were examined, a striking difference was observed between the two. For the construct lacking the *HIS3* terminator, normal *SER3* repression was observed, showing that the substitution of 150 bp of *SRG1* with *HIS3* did not impair repression (Fig. 3b, lane 2). By contrast, when the *HIS3* terminator was included, *SER3* repression was abolished (Fig. 3b, lane 3). Because the *SRG1* promoter is fully functional in the latter construct, but *SER3* is derepressed, we conclude that repression does not occur by promoter competition. These results also support the transcription-interference model, because termination of *SRG1* transcription 5' of the *SER3* UAS causes derepression of *SER3*.

Finally, we performed experiments to test directly the transcription-interference model. Because the direct activators of *SER3* have not been identified, we constructed strains in which the *SER3* UAS was replaced by Gal4-binding sites (Fig. 4a). In such strains we could assay the binding of a known transcription factor, Gal4. Our results show that in these constructs, *SER3* transcription is repressed even when cells are grown under inducing conditions for Gal4 activation, demonstrating that *SRG1*-dependent repression is intact (Fig. 4b, lane 2). By contrast, in an *srg1-1* mutant, *SER3* transcription is strongly galactose inducible (Fig. 4b, lane 4). The *GAL1* gene

serves as a control for galactose induction. We then used ChIP to measure the level of Gal4 bound *in vivo* to the *SER3* promoter in these constructs. Our results show that Gal4 binding is low when it is adjacent to wild-type *SRG1* (Fig. 4c, lanes 3 and 4; Supplementary Fig. 5a), but it is high when adjacent to *srg1-1* (Fig. 4c, lanes 7 and 8; Supplementary Fig. 5a). Thus, *SRG1* transcription inhibits Gal4 binding, providing direct support for a transcription-interference model.

As a second test of the interference model, we integrated the *SRG1* and *srg1-1* promoters at the *GAL7* locus, 5' of the *GAL7* UAS (Fig. 4d), and analysed both *GAL7* mRNA levels and the binding of Gal4. Transcription from the *SRG1* promoter, as detected by a *GAL7* 5' probe, appears to extend across *GAL7*, probably because the *SRG1* terminator is not present (Fig. 4e, lanes 3 and 4). Northern analysis shows that the *SRG1* promoter represses *GAL7* transcription (Fig. 4e, lane 4), consistent with the Gal⁻ phenotype observed for this strain. By contrast, the *srg1-1* mutant promoter has no detectable effect on *GAL7* expression (Fig. 4e, lane 6). ChIP analysis demonstrates that binding of Gal4 to the *GAL7* UAS is significantly reduced in the presence of the wild-type *SRG1* promoter but not in the presence of *srg1-1* (Fig. 4f, compare lanes 7 and 8 with lanes 11 and 12; Supplementary Fig. 5b). Thus, transcription from the *SRG1* promoter is sufficient to interfere with the binding of Gal4. Because only 31 bases of the *SRG1* transcript are included in this construct, these results also provide strong evidence against the RNA-mediated model. These results are similar to previous studies that demonstrated that interference with Gal4 binding at *GAL7* was caused by aberrant transcription readthrough from the *GAL10* gene²⁰.

Several aspects of *SRG1* repression remain to be elucidated,

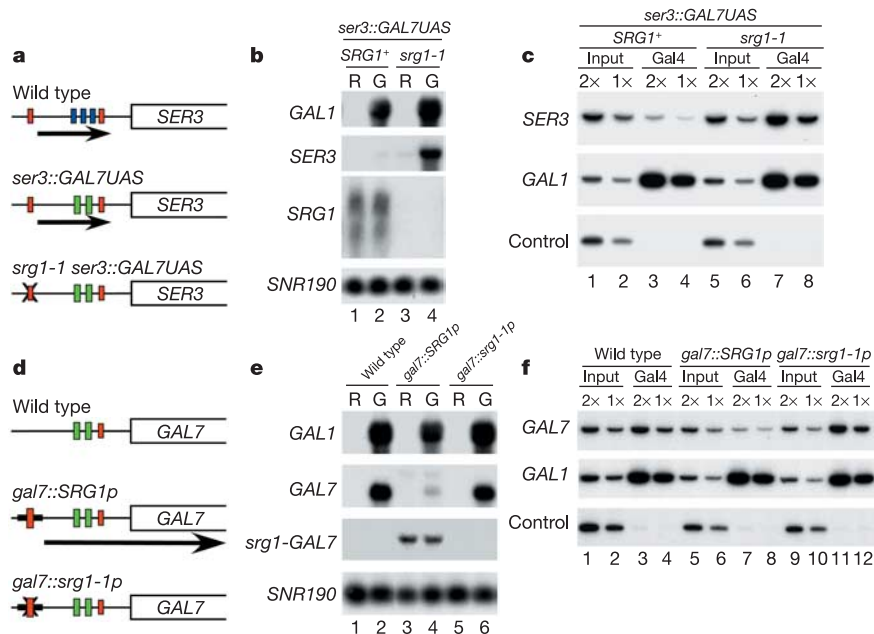


Figure 4 Two tests for transcription interference by *SRG1*. **a**, A diagram of a wild-type *SER3* allele (FY2097) and two *ser3::GAL7UAS* alleles in which the *GAL7UAS* region (with two Gal4-binding sites (green boxes)) has replaced the putative *SER3* UAS region (blue boxes) 3' of either the wild-type *SRG1* promoter (FY2259) or the *srg1-1* TATA mutant (FY2260). **b**, Northern analysis of *GAL1* (control for Gal4 activation), *SER3*, *SRG1* and *SNR190* (loading control) from the strains described in **a**. Strains were grown in media that is non-inducing (raffinose, R) and inducing (galactose, G) for Gal4 activity. **c**, ChIP analysis of Gal4 in the strains described in panel **a** that were grown in media containing galactose. Gal4 binding was measured at the Gal4-binding sites inserted at *SER3* (top panel), at *GAL1* as a positive control (middle panel) and at a region of chromosome V that lacks open reading frames as a negative control (bottom panel). Quantification of the

results is shown in Supplementary Fig. 5. **d**, A diagram of a wild-type *GAL7* allele (FY2256) and two *gal7* mutants in which either the wild-type *SRG1* promoter (FY2257; *gal7::SRG1p*) or the *srg1-1* TATA mutant promoter (FY2258; *gal7::srg1-1p*) is integrated 5' of the Gal4-binding sites (green boxes) in *GAL7*. **e**, Northern analysis of *GAL1* (a control for Gal4 activation), *GAL7*, *srg1-GAL7* and *SNR190* (loading control) from the strains described in panel **d**. Strains were grown in media that are non-inducing (raffinose, R) and inducing (galactose, G) for Gal4 activity. **f**, ChIP analysis of Gal4 binding at *GAL7* in the strains described in panel **d** that were grown in media containing galactose. Gal4 binding was measured at *GAL7* (top panel), at *GAL1* as a positive control (middle panel) and at a region of chromosome V that lacks open reading frames as a negative control (bottom panel). Quantification of the results is shown in Supplementary Fig. 5.

including its relationship with other *SER3* regulatory factors. Previous studies demonstrated that the Swi/Snf nucleosome-remodelling complex is also required for repression of *SER3*⁹. Thus, one possibility is that Swi/Snf activates *SRG1* transcription, conferring *SER3* repression. However, northern analysis (Supplementary Fig. 4) shows that *SRG1* RNA levels are only reduced two-fold in a *snf2Δ* mutant, indicating that the roles of Swi/Snf and *SRG1* in *SER3* repression may be independent. In addition, the expression and role of *SRG1* during *SER3* derepression by amino-acid starvation awaits further analysis.

Our studies have identified the first case of an intergenic transcript that represses by transcription interference. Given the large number of non-coding RNAs present in many organisms, including *Escherichia coli*, *S. cerevisiae* and humans^{1–6,24}, this mechanism of transcriptional control may be widespread. □

Methods

S. cerevisiae strains and media

All *S. cerevisiae* strains used (Supplementary Table 1) are isogenic with a *GAL2*⁺ derivative of S288C. Strains were constructed by standard methods, either by crosses or by transformation. Details are available upon request. The *ser3Δ::kanMX* allele has been previously described⁹. The *srg1-1* mutation changes the *SRG1* TATA sequence at position –558 (relative to the *SER3* ATG) from TATAAA to CCTAGG. The *srg1Δ1::HIS3* and *srg1Δ1::HIS3 + T* mutations replace *SRG1* sequence from –451 to –300 by either the *HIS3* ORF (+1 to +663 with +1 at the *HIS3* ATG) or the *HIS3* ORF and its transcription terminator (+1 to +817). The *ser3Δ::URA3* mutation replaces the *SER3*-coding sequence (+1 to +1410) with the coding sequence of *URA3* (+1 to +804). The *gal7::SRG1p* mutation inserts a sequence that includes the *SRG1* promoter and the two *SRG1* transcription-initiation sites (–713 to –445 relative to the *SER3* ATG) at position –378 of *GAL7*. Similar to *gal7::SRG1p*, the *gal7::srg1-1p* mutation inserts the *srg1-1* promoter sequence containing the mutant TATA sequence. The *ser3::GAL7UAS* mutation replaces *SER3* promoter sequences from –245 to –150 with *GAL7* sequences from –289 to –164, which includes two Gal4-binding sites. The *ser3Δ150::3MYC*, *ser3Δ250::3MYC* and *ser3Δ350::3MYC* mutations are internal deletions within *SRG1* that replace sequences from –660 to –150, –250 or –350 (relative to *SER3* ATG) with three copies of the Myc epitope tag. Unless otherwise noted, strains were grown in rich YPD medium, which is a repressing condition for *SER3* transcription.

Chromatin immunoprecipitation analysis

ChIP analysis was performed as previously described⁹. Antibodies used for immunoprecipitation were rabbit polyclonal anti-TBP serum (5 μl per immunoprecipitation; a gift of S. Buratowski), mouse monoclonal 12CA5 anti-HA ascites fluid (2 μl per immunoprecipitation; gift of E. Harlow), and rabbit polyclonal anti-Gal4(DD) serum (1 μl per immunoprecipitation; Santa Cruz Biotechnology). Dilutions of input DNA (1:500 and 1:1000) and immunoprecipitated DNA (1:5 and 1:10 for Rpb3-HA, 1:2.5 and 1:5 for TBP, and 1:10 and 1:20 for Gal4) were analysed by quantitative polymerase chain reaction (PCR) and the products were separated on a 6% non-denaturing polyacrylamide gel. The PCR primers amplify the following regions whose coordinates are given relative to the ATG (+1): *SER3* A primers amplify a 312-bp product from –1157 to –846; *SER3* B primers amplify a 315-bp product from –644 to –329; *SER3* C primers amplify a 298-bp product from –234 to +64; *SER3* D primers amplify a 318-bp fragment from +111 to +429; *GAL7* UAS primers amplify a 283-bp product from –370 to –88; and *ser3::GAL7* UAS primers amplify a 332-bp product from –454 to –123. The PCR primers used to amplify the *GAL1* UAS and the control region from chromosome V that lacks open reading frames have been previously described^{25,26}.

Transcription run-on analysis

The transcription run-on assay was performed as previously described²⁷, with the following modifications. YPD cultures of 200 ml were grown to approximately 3–5 × 10⁶ cells ml^{–1}. Transcription was allowed to proceed for 5 min at 30 °C after addition of 200 μCi of [α-³²P] rUTP. Total RNA was isolated using the hot phenol method⁹. After partial RNA hydrolysis with 0.2 M NaOH, labelled RNA was hybridized to oligonucleotides immobilized on GeneScreen (PerkinElmer). Membranes were washed with 2 × SSC/0.1% SDS for 30 min, twice at 37 °C and once at 42 °C. The nine probes used were antisense oligonucleotides corresponding to the following positions relative to the *SER3* ATG: –650 to –724 (1), –467 to –541 (2), –341 to –415 (3), –217 to –291 (4), –118 to –192 (5), +39 to –36 (6), +223 to +149 (7), +530 to +456 (8), +891 to +817 (9). See the gene map in Fig. 1c. Two sense oligonucleotides that correspond to positions –291 to –217 (N4) and –192 to –118 (N5) of *SER3* were used as negative controls. An antisense oligonucleotide corresponding to positions +437 to +363 of *ACT1* was included as a positive control.

RNA analysis

Northern hybridization analysis was performed as previously described⁹. Probes for *SER3*, *SRG1*, *SNR190*, *URA3*, *HIS3*, *GAL7*, *GAL7* 5' UTR, *GAL1* and *ACT1* were generated by randomly labelling PCR products that were amplified from genomic DNA. The coordinates of these probes relative to the ATG of each gene are as follows: *SER3* from +111 to +1342, *SRG1* from –454 to –123 (relative to *SER3* ATG), *URA3* from +206 to

+680, *HIS3* from –27 to +376, *GAL7* from +22 to +706, *GAL7* 5' UTR from –370 to –88, *GAL1* from +252 to +1329, and *ACT1* from –376 to +1015. Probes for *SNR190* and *TP11* have been previously described^{28,29}. Primer-extension analysis was performed as previously described³⁰. The *SER3* primer corresponds to bases +26 to +3 of the *SER3* coding sequence. The *SRG1* primer corresponds to bases –414 to –439 relative to the *SER3* ATG. The *ADH1* primer corresponds to bases +32 to +10 of the *ADH1*-coding sequence.

Received 2 February; accepted 2 April 2004; doi:10.1038/nature02538.

1. Cawley, S. *et al.* Unbiased mapping of transcription factor binding sites along human chromosomes 21 and 22 points to widespread regulation of noncoding RNAs. *Cell* **116**, 499–509 (2004).
2. Kapranov, P. *et al.* Large-scale transcriptional activity in chromosomes 21 and 22. *Science* **296**, 916–919 (2002).
3. Huhowitz, E. H. & Brown, P. O. Genome-wide analysis of mRNA lengths in *Saccharomyces cerevisiae*. *Genome Biol.* **5**, R2 (2003).
4. Rinn, J. L. *et al.* The transcriptional activity of human chromosome 22. *Genes Dev.* **17**, 529–540 (2003).
5. Saha, S. *et al.* Using the transcriptome to annotate the genome. *Nature Biotechnol.* **20**, 508–512 (2002).
6. Chen, J. *et al.* Identifying novel transcripts and novel genes in the human genome by using novel SAGE tags. *Proc. Natl Acad. Sci. USA* **99**, 12257–12262 (2002).
7. Gribnau, J., Diderich, K., Pruzina, S., Calzolari, R. & Fraser, P. Intergenic transcription and developmental remodeling of chromatin subdomains in the human β-globin locus. *Mol. Cell* **5**, 377–386 (2000).
8. Ashe, H. L., Monks, J., Wijgerde, M., Fraser, P. & Proudfoot, N. J. Intergenic transcription and transduction of the human β-globin locus. *Genes Dev.* **11**, 2494–2509 (1997).
9. Martens, J. A. & Winston, F. Evidence that Swi/Snf directly represses transcription in *S. cerevisiae*. *Genes Dev.* **16**, 2231–2236 (2002).
10. Albers, E., Laize, V., Blomberg, A., Hohmann, S. & Gustafsson, L. Ser3p (Yer081wp) and Ser33p (Yil074cp) are phosphoglycerate dehydrogenases in *Saccharomyces cerevisiae*. *J. Biol. Chem.* **278**, 10264–10272 (2003).
11. Cliften, P. *et al.* Finding functional features in *Saccharomyces* genomes by phylogenetic footprinting. *Science* **301**, 71–76 (2003).
12. Kellis, M., Patterson, N., Endrizzi, M., Birren, B. & Lander, E. S. Sequencing and comparison of yeast species to identify genes and regulatory elements. *Nature* **423**, 241–254 (2003).
13. van Hoof, A., Lennertz, P. & Parker, R. Yeast exosome mutants accumulate 3'-extended polyadenylated forms of U4 small nuclear RNA and small nucleolar RNAs. *Mol. Cell Biol.* **20**, 441–452 (2000).
14. Andersen, A. A. & Panning, B. Epigenetic gene regulation by noncoding RNAs. *Curr. Opin. Cell Biol.* **15**, 281–289 (2003).
15. Hirschman, J. E., Durbin, K. J. & Winston, F. Genetic evidence for promoter competition in *Saccharomyces cerevisiae*. *Mol. Cell Biol.* **8**, 4608–4615 (1988).
16. Emerman, M. & Temin, H. M. Genes with promoters in retrovirus vectors can be independently suppressed by an epigenetic mechanism. *Cell* **39**, 449–467 (1984).
17. Adhya, S. & Gottesman, M. Promoter occlusion: transcription through a promoter may inhibit its activity. *Cell* **29**, 939–944 (1982).
18. Corbin, V. & Maniatis, T. Role of transcriptional interference in the *Drosophila melanogaster* Adh promoter switch. *Nature* **337**, 279–282 (1989).
19. Cullen, B. R., Lomedico, P. T. & Ju, G. Transcriptional interference in avian retroviruses—implications for the promoter insertion model of leukaemogenesis. *Nature* **307**, 241–245 (1984).
20. Greger, I. H., Aranda, A. & Proudfoot, N. Balancing transcriptional interference and initiation on the *GAL7* promoter of *Saccharomyces cerevisiae*. *Proc. Natl Acad. Sci. USA* **97**, 8415–8420 (2000).
21. Hausler, B. & Somerville, R. L. Interaction in vivo between strong closely spaced constitutive promoters. *J. Mol. Biol.* **127**, 353–356 (1979).
22. Proudfoot, N. J. Transcriptional interference and termination between duplicated alpha-globin gene constructs suggests a novel mechanism for gene regulation. *Nature* **322**, 562–565 (1986).
23. Valerius, O., Brendel, C., Duvel, K. & Braus, G. H. Multiple factors prevent transcriptional interference at the yeast *ARO4-HIS7* locus. *J. Biol. Chem.* **277**, 21440–21445 (2002).
24. Gottesman, S. Stealth regulation: biological circuits with small RNA switches. *Genes Dev.* **16**, 2829–2842 (2002).
25. Komarnitsky, P., Cho, E. J. & Buratowski, S. Different phosphorylated forms of RNA polymerase II and associated mRNA processing factors during transcription. *Genes Dev.* **14**, 2452–2460 (2000).
26. Dudley, A. M., Rougeulle, C. & Winston, F. The Spt components of SAGA facilitate TBP binding to a promoter at a post-activator-binding step in vivo. *Genes Dev.* **13**, 2940–2945 (1999).
27. Birse, C. E., Lee, B. A., Hansen, K. & Proudfoot, N. J. Transcriptional termination signals for RNA polymerase II in fission yeast. *EMBO J.* **16**, 3633–3643 (1997).
28. Duina, A. A. & Winston, F. Analysis of a mutant histone H3 that perturbs the association of Swi/Snf with chromatin. *Mol. Cell Biol.* **24**, 561–572 (2004).
29. Hirschhorn, J. N., Brown, S. A., Clark, C. D. & Winston, F. Evidence that SNF2/SWI2 and SNF5 activate transcription in yeast by altering chromatin structure. *Genes Dev.* **6**, 2288–2298 (1992).
30. Dudley, A. M., Gansheroff, L. J. & Winston, F. Specific components of the SAGA complex are required for Gcn4- and Gcr1-mediated activation of the *his4-912d* promoter in *Saccharomyces cerevisiae*. *Genetics* **151**, 1365–1378 (1999).

Supplementary Information accompanies this paper on www.nature.com/nature.

Acknowledgements We thank A. Dudley, A. Duina and J. Wu for comments on the manuscript. We also thank P. Cliften, P. Sudarsanam and M. Johnston for sharing unpublished results, and J. O'Sullivan and N. Proudfoot for advice on transcription run-on experiments. This work was supported by a grant from the National Institutes of Health to F.W. and by a postdoctoral fellowship from the Canadian Institute for Health Research to J.A.M.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to F.W. (winston@genetics.med.harvard.edu).

Nonsense-mediated messenger RNA decay is initiated by endonucleolytic cleavage in *Drosophila*

David Gatfield & Elisa Izaurralde

European Molecular Biology Laboratory, Meyerhofstrasse 1, D-69117 Heidelberg, Germany

In eukaryotic cells, messenger RNAs harbouring premature termination codons (PTCs) are rapidly degraded by a conserved post-transcriptional mechanism referred to as nonsense-mediated mRNA decay (NMD)^{1,2}, which prevents the synthesis of truncated proteins that could be deleterious for the cell^{1,2}. Studies in yeast and mammals indicate that degradation by means of this pathway can occur from both the 5' end of the message (involving decapping and 5'-to-3' exonucleolytic digestion by XRN1) or the 3' end (through accelerated deadenylation and exosome-mediated 3'-to-5' decay)³⁻⁹. Here we show that, contrary to expectation, degradation of PTC-containing messages in *Drosophila* is initiated by endonucleolytic cleavage(s) in the vicinity of the nonsense codon. The resulting 5' fragment is rapidly degraded by exonucleolytic digestion by the exosome, whereas the 3' fragment is degraded by XRN1. This decay route is shown for several PTC-containing reporters, as well as an endogenous mRNA that is naturally regulated by NMD. We conclude that, despite conservation in the NMD machinery, PTC-containing transcripts are degraded in *Drosophila* by a mechanism that differs considerably from those described in yeast and mammals^{3,6,7}.

In eukaryotic cells, recognition of a PTC by translating ribosomes leads to the assembly of a so-called 'surveillance complex' on the mRNA, consisting minimally of the proteins UPF1, UPF2 and UPF3, which targets the transcript for degradation^{1,2}. Using *Drosophila* Schneider cell lines expressing reporters harbouring PTCs, we sought to characterize the nucleolytic activities contributing to the decay of nonsense-containing messages in this organism. To this end, we silenced with specific double-stranded RNAs (dsRNAs) the expression of factors involved in decapping (DCP1, DCP2, LSM1), 5'-to-3' degradation (XRN1, RAT1), deadenylation (CCR4, CAF1, PAN2, PAN3) and 3'-to-5' decay (the exosome components CSL4, RRP4, RRP6 and the helicase SKI2)^{1,2}, both singly and in combinations, but in no case were we able to detect a significant stabilization of the full-length PTC-containing reporter (data not shown).

Surprisingly, depletion of XRN1 led to the appearance of a stable mRNA species with higher electrophoretic mobility (Fig. 1a-c, arrowheads). This shorter RNA species was never observed for the wild-type reporters but was readily detected for all three PTC-containing reporters tested, which derived from genes encoding *Drosophila* alcohol dehydrogenase (adh; Fig. 1a), bacterial chloramphenicol acetyl transferase (CAT; Fig. 1b) or green fluorescent protein (GFP; Fig. 1c). Simultaneous depletion of XRN1 and UPF1 prevented the appearance of this species, indicating that it was a decay intermediate arising from the messages undergoing NMD (Fig. 1a, lane 12).

The decay intermediate was detected with a probe that hybridized to a region of the adh reporter downstream, but not with a probe that hybridized upstream of the PTC (Fig. 1d). We therefore conclude that a 5'-truncated degradation intermediate accumulates in XRN1-depleted cells. In *Saccharomyces cerevisiae*, deletion of the XRN1 gene also leads to the accumulation of 5'-truncated PTC-containing messages¹⁰. However, the nature of these decay intermediates differs substantially from those seen in *Drosophila*: in

yeast, messages are truncated by only a few nucleotides, and inactivation of Rat1p prevents their formation¹⁰. The *Drosophila* intermediates are several hundred nucleotides shorter than the full-length reporter (Fig. 1d), and simultaneously depleting RAT1 had no effect on their appearance (Fig. 1a, b).

Although we observed the decay intermediate for three different

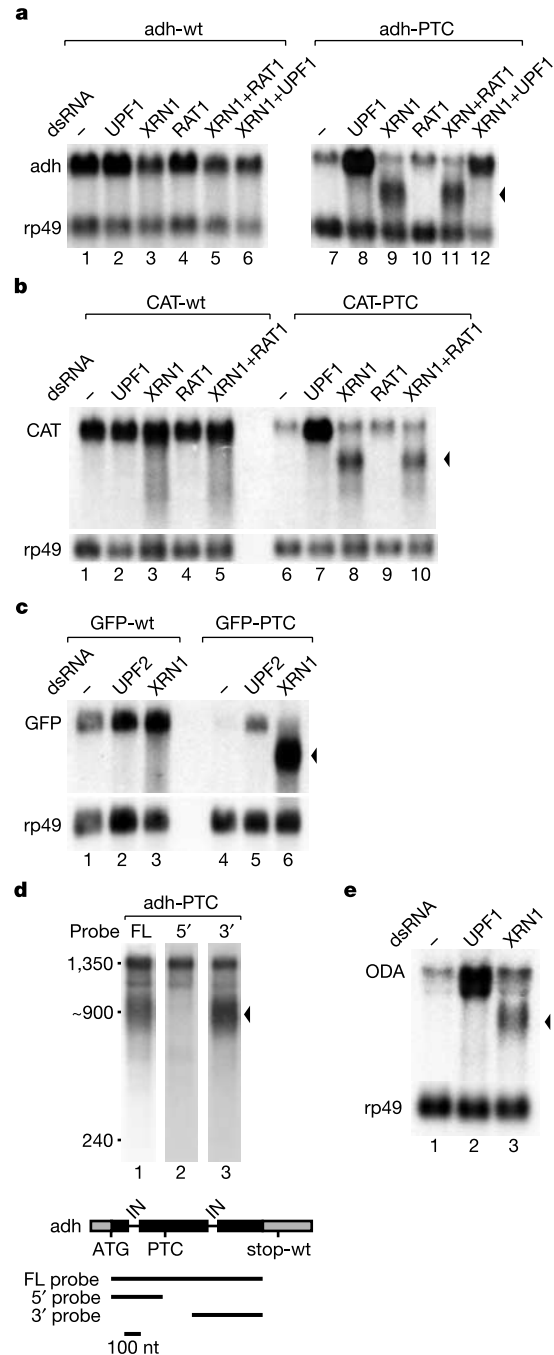


Figure 1 Accumulation of a 5'-truncated intermediate in XRN1-depleted cells. **a-c**, Steady-state levels of wild-type (wt) or PTC-containing reporters derived from adh (**a**), CAT (**b**) or GFP (**c**), in control cells or cells treated with the dsRNAs indicated above the lanes. rp49 mRNA served as a loading control. Arrowheads indicate the position of the decay intermediate. **d**, The steady-state level of adh-PTC in XRN1-depleted cells was analysed by using 5' or 3' probes as illustrated below the panel. Black and grey boxes in the diagram of the reporter represent adh exons and vector sequences, respectively. IN, introns; nt, nucleotides. **e**, Steady-state levels of endogenous ODA mRNA in control and depleted cells.

reporters, it was important to determine whether an endogenous NMD substrate followed the same degradation pathway. With a microarray analysis of total RNA preparations from Schneider cells depleted of NMD factors (J. Rehwinkel, D.G. and E.I., unpublished observations), we identified the ornithine decarboxylase antizyme (ODA) transcript¹¹ as an endogenous message regulated by NMD. Consistently, depletion of UPF1 led to a strong increase in the ODA mRNA expression levels (Fig. 1e). Analysis of the ODA transcript in XRN1-depleted cells revealed the presence of a degradation intermediate (Fig. 1e), indicating that this endogenous NMD substrate followed the same decay pathway as the reporters.

To determine whether the length of the intermediate was related to the position of the PTC, we introduced stop codons over the whole length of the coding region of the *adh* reporter (Fig. 2a). When depleting XRN1 in the corresponding stably transfected cells, we noticed a correlation between the size of the decay intermediate and the position of the nonsense codon (Fig. 2b). The length of the decay intermediate is therefore determined by the position at which the PTC is inserted. We envisaged several possibilities to explain this

position-dependent effect. First, the nonsense message might be degraded from the 5' end by an unknown exonuclease, which ceases degradation near the PTC (for example, because of inhibition by the stalled ribosome), and degradation of the remaining 3' fragment requires XRN1. Second, the residual XRN1 molecules in depleted cells might be recruited less efficiently (or might be less processive) so that their progression along the mRNA is blocked by the stalled ribosome. Last, an endonuclease might initiate decay by cleaving the mRNA at a position dependent on the location of the PTC. The resulting 3' fragment would be degraded by XRN1, whereas the 5' fragment could potentially be the substrate of the 3'-to-5' exonucleases of the exosome.

If the decay of PTC-containing transcripts is initiated by endonucleolytic cleavage, a 5' fragment should be generated. However, it is likely that this fragment is degraded very rapidly, because we did

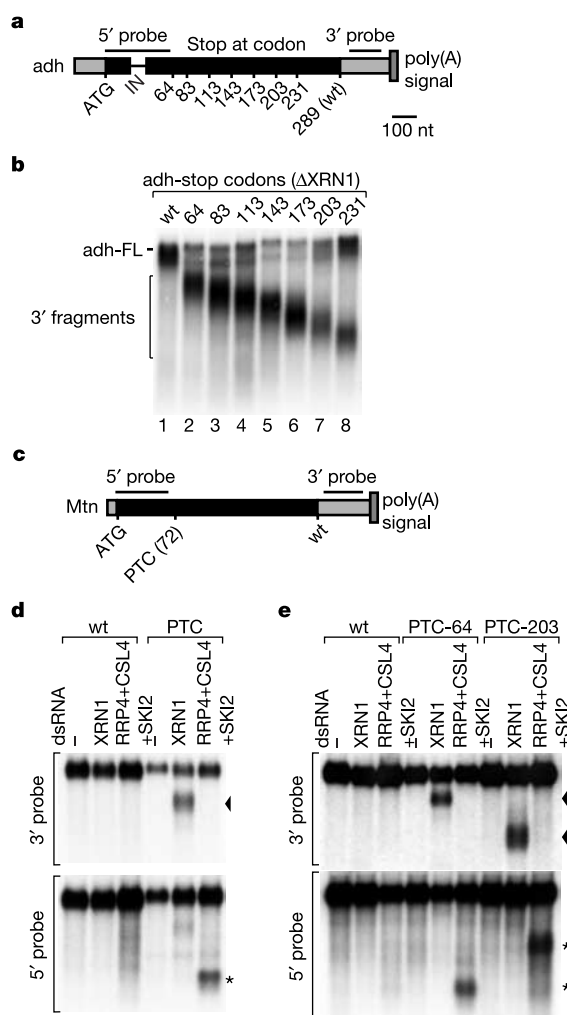


Figure 2 The size of the decay intermediates is correlated with the position of the PTC. **a**, Diagram of the *adh* reporter with stop codons inserted at different positions. Symbols are as in Fig. 1d. nt, nucleotides. **b**, Steady-state levels of *adh* wild-type (*adh*-wt) or *adh*-PTC reporters in XRN1-depleted cells. FL, full-length. **c**, Schematic representation of the inducible *CAT* reporter. Black boxes, *CAT* cDNA; grey boxes, vector sequences; *Mtn*, metallothionein promoter. **d**, **e**, Steady-state levels of *CAT* (**d**) or *adh* (**e**) reporters in cells treated with the indicated dsRNAs. Northern blots were hybridized with the 5' or 3' probes shown in **a** and **b**. Arrowheads and asterisks indicate the position of the 3' or 5' intermediates, respectively.

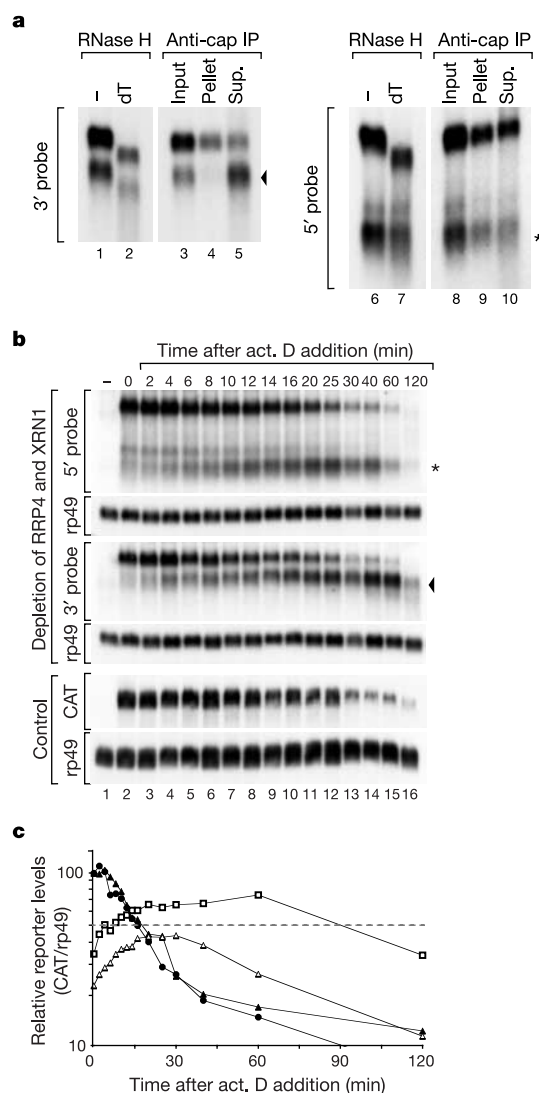


Figure 3 The 5' intermediate is capped, whereas the 3' intermediate is polyadenylated. **a**, Cells expressing inducible *CAT*-PTC were depleted of both RRP4 and XRN1. In lanes 1, 2, 6 and 7, total RNA samples were subjected to oligo(dT)-targeted RNase H cleavage. In lanes 3–5 and 8–10, total RNA samples were immunoprecipitated (IP) with anti-cap antibodies. Northern blots were hybridized with 5' or 3' probes as indicated. Arrowheads and asterisks indicate the position of the 3' and 5' intermediates, respectively. **b**, Decay rates of inducible *CAT* transcripts in control cells or cells depleted of both RRP4 and XRN1. act. D, actinomycin D. **c**, The levels of *CAT* transcripts relative to rp49 mRNA in **b** are plotted as a function of time: open triangles, 5' fragment; squares, 3' fragment; filled triangles, full length in control cells; circles, full length in depleted cells.

not detect a 5' fragment in cell lines expressing the NMD reporters constitutively, despite the depletion of exosome components (data not shown). We therefore expressed the NMD reporters under the control of a strong inducible promoter (metallothionein). We expected that overexpression of the reporter might saturate the mRNA decay machinery. In addition, short induction times should yield a more homogeneous population of decay intermediates, making them more readily detectable.

The reporters under the control of the metallothionein promoter were subject to NMD, as shown by measuring decay rates and by the presence of the characteristic intermediate in XRN1-depleted cells (Supplementary Fig. 1). Simultaneous depletion of the exosome components RRP4 and CSL4 and of SKI2, followed by induction of the reporters (namely CAT or adh), led to the appearance of an mRNA fragment detected with a probe hybridizing to a region of the reporters 5' to the PTC (Fig. 2c, d, e, asterisks). Similar results

were obtained when exosome subunits were depleted individually (Supplementary Fig. 2). Remarkably, the length of the 5' fragment is also determined by the position at which the PTC is inserted (Fig. 2e).

For fragments generated by endonucleolytic cleavage, one would expect the 5' fragment to possess a cap structure and the 3' fragment a poly(A) tail. Indeed, the 3' fragment changed its mobility after removal of the poly(A) tail by oligo(dT)-targeted RNase H cleavage (Fig. 3a, lanes 1 and 2), indicating that it is polyadenylated. Moreover, the 5' fragment, but not the 3' fragment, could be immunoprecipitated together with antibodies recognizing the cap structure (Fig. 3a, lanes 9 and 4). Similar results were obtained for the intermediates derived from adh-64 and adh-203 (Supplementary Fig. 2, and data not shown).

To demonstrate a precursor-product relationship between the full-length mRNA and the decay intermediates, the half-lives of inducible CAT-PTC transcripts were determined after a short induction pulse followed by the inhibition of transcription by actinomycin D. In control cells, the half-life of CAT-PTC was about 15 min. In cells depleted of both RRP4 and XRN1, the half-life of CAT-PTC was unchanged, indicating that the first step(s) in the decay pathway of this mRNA (for example, endonucleolytic cleavage) were not significantly affected by the depletions (Fig. 3b, c). The appearance of the intermediates was inversely correlated with the decay of the full-length transcript, indicating that the intermediates are generated from the full-length mRNA undergoing NMD (Fig. 3b, c). The efficient and rapid conversion of the full-length transcript to its fragments underscores the observation that the decay mechanism uncovered by XRN1 and exosome depletion is the major NMD pathway in *Drosophila* and is unlikely to be a non-specific effect of treatment with dsRNA¹².

By S1 nuclease protection assays and primer extension, we mapped the ends of the 5' and 3' fragments more precisely for CAT-PTC in cells depleted of both RRP4 and XRN1 (Fig. 4). Analysis of the 3' intermediate by primer extension revealed five major products corresponding to cleavages centred 23 and 10 nucleotides upstream of the PTC (positions -23 and -10) or 11–18, 23 and 71–83 nucleotides downstream of the PTC (positions +[11–18], +23 and +[71–83]; Fig. 4a, b). The cleavages at positions -23, -10, +18, +25 and +[76–86] were confirmed by S1 nuclease protection mapping (Fig. 4c). S1 nuclease protection assay to map the 3' end of the 5' intermediate led to protected fragments corresponding to cleavages centred at positions +11 and +[64–80] (Fig. 4d). The protected fragments were not observed in control cells or for the wild-type reporter in depleted cells (Fig. 4b–d).

The cleavage sites were also determined for adh-64, adh-203 and ODA mRNAs (Supplementary Fig. 3, and data not shown). For the different reporters, the sites varied in distance to the PTCs, suggesting that cleavage does not occur by a strict 'molecular ruler' mechanism and might depend on nucleotide sequences. In all cases, the cleavage products were in the vicinity of the PTC and were heterogeneous. This heterogeneity might reflect the trimming of the free 5' or 3' ends of the intermediates by residual amounts of XRN1 or RRP4, respectively, which were depleted to about 10% of wild-type levels (Supplementary Figs 1 and 2).

Studies in yeast and mammals have led to a model in which the degradation of mRNAs harbouring PTCs occurs from both ends of the transcript^{3,6,7}. However, these models cannot account for the decay intermediates observed in *Drosophila*. Indeed, the accumulation of a capped 5' fragment in cells depleted of exosome subunits, despite the presence of the decapping enzymes and XRN1, is unexpected according to current models^{3,6,7} and indicates that NMD substrates in *Drosophila* are unlikely to be degraded by decapping followed by 5'-to-3' exonucleolytic digestion by XRN1. Conversely, the stabilization of a polyadenylated 3' fragment in XRN1-depleted cells despite the presence of deadenylases and the exosome suggests that accelerated deadenylation, followed by

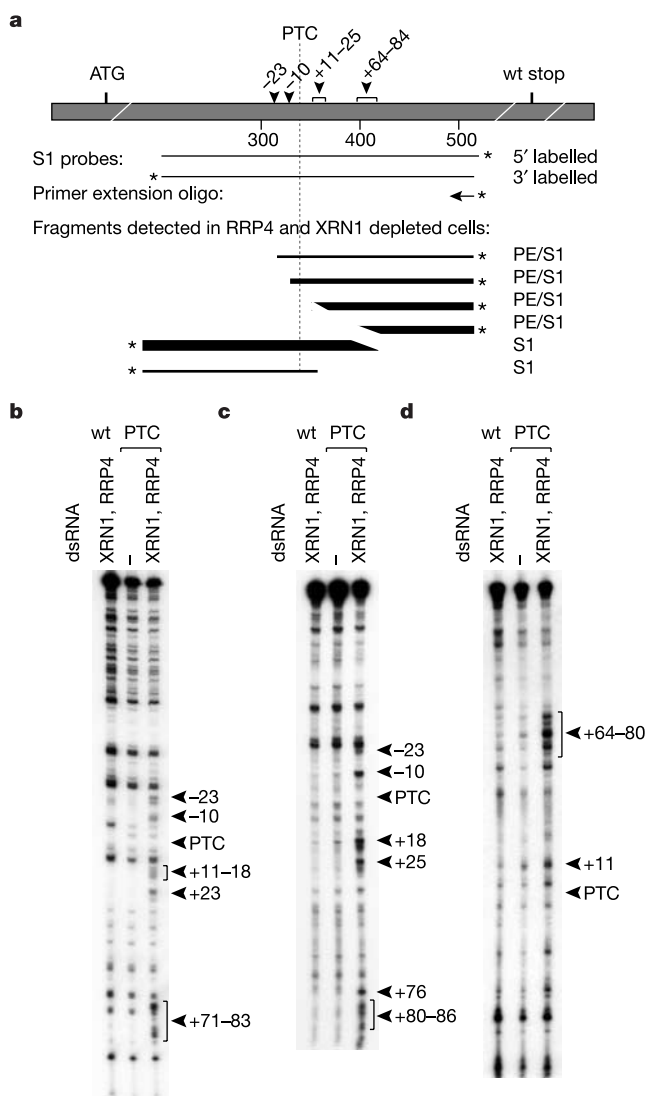


Figure 4 Endonucleolytic cleavage(s) of nonsense transcripts occurs in the vicinity of the PTC. **a**, Diagram of the CAT reporter near the PTC. The positions of the S1 probes and the oligonucleotide used for primer extension, and also the major products from the S1 protections (S1) and primer extensions (PE), are shown. Scale bar, nucleotides. **b–d**, Cells expressing inducible CAT wild-type (CAT-wt) or CAT-PTC were depleted of both RRP4 and XRN1. RNA samples were analysed by primer extension (**b**) or S1 nuclease mapping with the probe 5' labelled (**c**) or 3' labelled (**d**). The major cleavage products relative to the PTC are shown. These products were not detected for CAT-PTC in control cells or for CAT-wt in depleted cells.

degradation of the mRNA body from the 3' end by the exosome^{3,6,7}, is unlikely to be a major pathway in *Drosophila*. Alternative models, wherein the stalled ribosome blocks the progression of XRN1 or of the exosome along the RNA are also unlikely: first, for these models to account for the observed intermediates, the two decay pathways should be mutually exclusive; second, the half-life of the full-length PTC-containing transcript remains unchanged in RRP4 and XRN1 co-depleted cells, arguing against a reduced processivity of the enzymes. The simplest model accounting for the results described above is that the decay intermediates originate from one or more endonucleolytic cleavages in the vicinity of the PTC.

Although most mRNA degradation occurs by exonucleolytic activities from both the 5' and the 3' end of the transcript², a growing number of mRNAs have been found to be degraded through endonucleolytic cleavage^{13–25}. Moreover, in metazoans the decay of mRNAs targeted by RNA-mediated interference is initiated by endonucleolytic cleavage²⁶; in *Escherichia coli*, stalling of the ribosome at a stop codon triggers the endonucleolytic cleavage of the mRNA at or in the vicinity of the stop^{27,28}. Here we show that endonucleolytic cleavage in the vicinity of the PTC is the predominant NMD pathway in *Drosophila* cells. It will be of interest to see whether this pathway makes a contribution to the decay of PTC-containing messages in other organisms. □

Methods

Cell culture and RNA interference

The establishment of stable cell lines expressing the NMD reporters and RNA-mediated interference were performed as described²⁹; dsRNAs corresponding to *Drosophila* XRN1 (*pcm*, gene CG3291 in FlyBase (<http://flybase.net/>), RAT1 (CG10354), RRP4 (CG3931), CSL4 (CG6249) and SKI2 (*tsr*, CG10210) were transcribed *in vitro* from cDNA fragments about 700 base pairs long amplified from a Schneider cell cDNA library.

Construction of NMD reporters

The NMD reporters adh-64 (nBR114), CAT-72, GFP-52 (used in Fig. 1) and adhΔIN3 (Fig. 2a, b) have been described previously²⁹; additional PTCs were introduced in the adhΔIN3 wild-type reporter using an oligonucleotide-directed *in vitro* mutagenesis system (quick-change site-directed mutagenesis; Stratagene). The presence of the stop codon was confirmed by sequencing. To generate the metallothionein-inducible reporters, a *KpnI/SalI* (blunt) fragment comprising the whole reporter sequence and the polyadenylation site was excised from the adh and CAT reporters described previously and cloned into the *KpnI/XhoI* (blunt) sites of vector pRmHa³⁰ containing the metallothionein promoter.

RNA analysis

RNA preparation and northern blot analysis were performed essentially as described²⁹. RNAs from inducible cell lines were harvested 180 min after induction with 1 mM copper sulphate. For the determination of RNA half-lives, control or depleted cells were incubated with actinomycin D (5 μg ml⁻¹) for the durations indicated in the figures. Total RNA samples were analysed by northern blotting. The amounts of the reporters and intermediates were quantified and normalized to rp49 mRNA (whose amounts relative to 18S rRNA do not change within the time frame of the experiments).

Body-labelled DNA probes were generated by unidirectional polymerase chain reaction (PCR) in accordance with standard protocols. A template for the ODA probe (encompassing nucleotides 168–898 of sequence accession no. NM_165863 GenBank) was amplified from a Schneider cell cDNA library. Primer extension analysis and S1 nuclease protection assays were performed in accordance with standard protocols on total RNA preparations from Schneider cells depleted of both RRP4 and XRN1. A single-stranded 5'-labelled probe complementary to nucleotides 189–504 of the CAT reporter was prepared by unidirectional PCR with a ³²P-labelled antisense oligonucleotide (5'-GGTTTTTCACCGTAACACGCCACATC-3'). To map the 3' end of the 5' fragment, a single-stranded probe complementary to nucleotides 188–504 of the CAT reporter was generated by unidirectional PCR, followed by tailing with [³²P]dideoxyATP with the use of terminal transferase (Roche). S1 protection for the 5' fragment was performed on poly(A)⁺ RNA. The S1-resistant fragments and primer extension products were analysed by electrophoresis on 4% sequencing gels. RNase H (USB, Cleveland, Ohio) digestion with a (dT)₁₅ oligonucleotide was performed in accordance with the manufacturer's instructions.

Immunoprecipitation of capped mRNAs was performed from total RNA preparations with the monoclonal antibody H20 in NET buffer (50 mM Tris-HCl pH 7.5, 150 mM NaCl, 0.1% Nonidet P40, 1 mM EDTA). One-fifth of the inputs and 100% of the immunoprecipitates were analysed by northern blotting.

Received 19 March; accepted 14 April 2004; doi:10.1038/nature02559.

- Wagner, E. & Lykke-Andersen, J. mRNA surveillance: the perfect persist. *J. Cell Sci.* **115**, 3033–3038 (2002).
- Wilusz, C. J., Wormington, M., Stuart, W. & Peltz, S. W. The cap-to-tail guide to mRNA turnover. *Nature Rev. Mol. Cell Biol.* **2**, 237–246 (2001).

- Cao, D. & Parker, R. Computational modeling and experimental analysis of nonsense-mediated decay in yeast. *Cell* **113**, 533–545 (2003).
- Chen, C. Y. & Shyu, A. B. Rapid deadenylation triggered by a nonsense codon precedes decay of the RNA body in a mammalian cytoplasmic nonsense-mediated decay pathway. *Mol. Cell. Biol.* **23**, 4805–4813 (2003).
- Hagan, K. W., Ruiz-Echevarria, M. J., Quan, Y. & Peltz, S. W. Characterization of cis-acting sequences and decay intermediates involved in nonsense-mediated mRNA turnover. *Mol. Cell. Biol.* **15**, 809–823 (1995).
- Lejeune, F., Li, X. & Maquat, L. E. Nonsense-mediated mRNA decay in mammalian cells involves decapping, deadenylation, and exonucleolytic activities. *Mol. Cell* **12**, 675–687 (2003).
- Mitchell, P. & Tollervey, D. An NMD pathway in yeast involving accelerated deadenylation and exosome-mediated 3' → 5' degradation. *Mol. Cell* **11**, 1405–1413 (2003).
- Muhlrad, D. & Parker, R. Premature translational termination triggers mRNA decapping. *Nature* **370**, 578–581 (1994).
- Muhlrad, D. & Parker, R. Aberrant mRNAs with extended 3' UTRs are substrates for rapid degradation by mRNA surveillance. *RNA* **5**, 1299–1307 (1999).
- He, F. & Jacobson, A. Upf1p, Nmd2p, and Upf3p regulate the decapping and exonucleolytic degradation of both nonsense-containing mRNAs and wild-type mRNAs. *Mol. Cell. Biol.* **21**, 1515–1530 (2001).
- Ivanov, I. P., Simin, K., Letsou, A., Atkins, J. F. & Gesteland, R. F. The *Drosophila* gene for antizyme requires ribosomal frameshifting for expression and contains an intronic gene for snRNP Sm D3 on the opposite strand. *Mol. Cell. Biol.* **18**, 1553–1561 (1998).
- Persengiev, S. P., Zhu, X. & Green, M. R. Nonspecific, concentration-dependent stimulation and repression of mammalian gene expression by small interfering RNAs (siRNAs). *RNA* **10**, 12–18 (2004).
- Binder, R., Hwang, S. P., Ratnasabapathy, R. & Williams, D. L. Degradation of apolipoprotein II mRNA occurs via endonucleolytic cleavage at 5'-AAU-3'/5'-UAA-3' elements in single-stranded loop domains of the 3'-noncoding region. *J. Biol. Chem.* **264**, 16910–16918 (1989).
- Binder, R. *et al.* Evidence that the pathway of transferin receptor mRNA degradation involves an endonucleolytic cleavage within the 3' UTR and does not involve poly(A) tail shortening. *EMBO J.* **13**, 1969–1980 (1994).
- Bremer, K. A., Stevens, A. & Schoenberg, D. R. An endonuclease activity similar to *Xenopus* PMR1 catalyzes the degradation of normal and nonsense-containing human beta-globin mRNA in erythroid cells. *RNA* **9**, 1157–1167 (2003).
- Brown, B. D. & Harland, R. M. Endonucleolytic cleavage of a maternal homeo box mRNA in *Xenopus* oocytes. *Genes Dev.* **4**, 1925–1935 (1990).
- Chernokalskaya, E. *et al.* A polysomal ribonuclease involved in the destabilization of albumin mRNA is a novel member of the peroxidase gene family. *RNA* **4**, 1537–1548 (1998).
- Cunningham, K. S., Dodson, R. E., Nagel, M. A., Shapiro, D. J. & Schoenberg, D. R. Vigilin binding selectively inhibits cleavage of the vitellogenin mRNA 3'-untranslated region by the mRNA endonuclease polysomal ribonuclease 1. *Proc. Natl Acad. Sci. USA* **97**, 12498–12502 (2000).
- Gallouzi, I. E. *et al.* A novel phosphorylation-dependent RNase activity of GAP-SH3 binding protein: a potential link between signal transduction and RNA stability. *Mol. Cell. Biol.* **18**, 3956–3965 (1998).
- Lee, C. H., Leeds, P. & Ross, J. Purification and characterization of a polysome-associated endoribonuclease that degrades c-myc mRNA *in vitro*. *J. Biol. Chem.* **273**, 25261–25271 (1998).
- Lim, S. K. & Maquat, L. E. Human beta-globin mRNAs that harbor a nonsense codon are degraded in murine erythroid tissues to intermediates lacking regions of exon I or exons I and II that have a cap-like structure at the 5' termini. *EMBO J.* **11**, 3271–3278 (1992).
- Scheper, W., Meinsma, D., Holthuisen, P. E. & Sussenbach, J. S. Long-range RNA interaction of two sequence elements required for endonucleolytic cleavage of human insulin-like growth factor II mRNAs. *Mol. Cell. Biol.* **15**, 235–245 (1995).
- Stevens, A. *et al.* β-Globin mRNA decay in erythroid cells: UG site-preferred endonucleolytic cleavage that is augmented by a premature termination codon. *Proc. Natl Acad. Sci. USA* **99**, 12741–12746 (2002).
- van Dijk, E. L., Sussenbach, J. S. & Holthuisen, P. E. Identification of RNA sequences and structures involved in site-specific cleavage of IGF-II mRNAs. *RNA* **4**, 1623–1635 (1998).
- Wang, Z. & Kiledjian, M. Identification of an erythroid-enriched endoribonuclease activity involved in specific mRNA cleavage. *EMBO J.* **19**, 295–305 (2000).
- Dykxhoorn, D. M., Novina, C. D. & Sharp, P. A. Killing the messenger: short RNAs that silence gene expression. *Nature Rev. Mol. Cell Biol.* **4**, 457–467 (2003).
- Hayes, C. S. & Sauer, R. T. Cleavage of the A site mRNA codon during ribosome pausing provides a mechanism for translational quality control. *Mol. Cell* **12**, 903–911 (2003).
- Sunohara, T., Jojima, K., Yamamoto, Y., Inada, T. & Aiba, H. Nascent-peptide-mediated ribosome stalling at a stop codon induces mRNA cleavage resulting in nonstop mRNA that is recognized by tmRNA. *RNA* **10**, 378–386 (2004).
- Gatfield, D., Unterholzner, L., Ciccarelli, F. D., Bork, P. & Izaurralde, E. Nonsense-mediated mRNA decay in *Drosophila*: at the intersection of the yeast and mammalian pathways. *EMBO J.* **22**, 3960–3970 (2003).
- Bunch, T. A., Grinblat, Y. & Goldstein, L. S. Characterization and use of the *Drosophila* metallothionein promoter in cultured *Drosophila melanogaster* cells. *Nucleic Acids Res.* **16**, 1043–1061 (1988).

Supplementary Information accompanies the paper on www.nature.com/nature.

Acknowledgements We thank J. Rehwinkel for sharing information on the microarray analysis, K. Czaplinski for helpful discussions, and E. D. Andriulis, J. T. Lis, R. Lührmann and S. F. Newbury for antibodies. This study was supported by the European Molecular Biology Organisation (EMBO) and the Human Frontier Science Program Organization (HFSPRO).

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to E.I. (izaurralde@embl-heidelberg.de).

Cyclic electron flow around photosystem I is essential for photosynthesis

Yuri Munekage^{1,2*}, Mihoko Hashimoto¹, Chikahiro Miyake², Ken-ichi Tomizawa², Tsuyoshi Endo³, Masao Tasaka¹ & Toshiharu Shikanai^{1*}

¹Graduate School of Biological Sciences, Nara Institute of Science and Technology, Ikoma, Nara 630-0101, Japan

²Plant Research Group, Research Institute of Innovative Technology for the Earth, Kizu, Kyoto 619-0225, Japan

³Graduate School of Biostudies, Kyoto University, Sakyo-ku, Kyoto 606-8502, Japan

* Present addresses: CEA Cadarache, Department of Plant Ecophysiology and Microbiology, 13108 Saint-Paul-lez-Durance Cedex, France (Y.M.); Graduate School of Agriculture, Kyushu University, Higashiku, Fukuoka 812-8581, Japan (T.S.)

Photosynthesis provides at least two routes through which light energy can be used to generate a proton gradient across the thylakoid membrane of chloroplasts, which is subsequently used to synthesize ATP. In the first route, electrons released from water in photosystem II (PSII) are eventually transferred to NADP⁺ by way of photosystem I (PSI)¹. This linear electron flow is driven by two photochemical reactions that function in series. The cytochrome *b₆f* complex mediates electron transport between the two photosystems and generates the proton gradient (Δ pH). In the second route, driven solely by PSI, electrons can be recycled from

either reduced ferredoxin or NADPH to plastoquinone, and subsequently to the cytochrome *b₆f* complex^{2–5}. Such cyclic flow generates Δ pH and thus ATP without the accumulation of reduced species. Whereas linear flow from water to NADP⁺ is commonly used to explain the function of the light-dependent reactions of photosynthesis, the role of cyclic flow is less clear. In higher plants cyclic flow consists of two partially redundant pathways. Here we have constructed mutants in *Arabidopsis thaliana* in which both PSI cyclic pathways are impaired, and present evidence that cyclic flow is essential for efficient photosynthesis.

In addition to ATP synthesis, generation of Δ pH is required for plants to cope with the changing quantities of light available under natural conditions. Whenever absorption of light energy exceeds its use, increased acidification of the thylakoid lumen induces thermal dissipation of absorbed energy from PSII antennae⁶. An *Arabidopsis* mutant, *pgr5* (proton gradient regulation), was identified by its high chlorophyll fluorescence due to a reduced level of thermal dissipation⁷. *pgr5* mutants are defective in cyclic flow, indicating that cyclic flow is essential for increased acidification of the thylakoid lumen and the induction of thermal dissipation.

The chloroplast genome of higher plants contains *ndh* genes encoding 11 subunits of the chloroplast NDH complex (NAD(P)H dehydrogenase). This complex was identified as a homologue of the mitochondrial complex I (NADH dehydrogenase)⁸. Whereas PGR5-dependent cyclic flow corresponds to the commonly known pathway in higher plants mediated by ferredoxin–plastoquinone reductase^{9–11}, the NDH-dependent pathway was first identified in cyanobacteria^{12,13}. Knockout of tobacco *ndh* genes

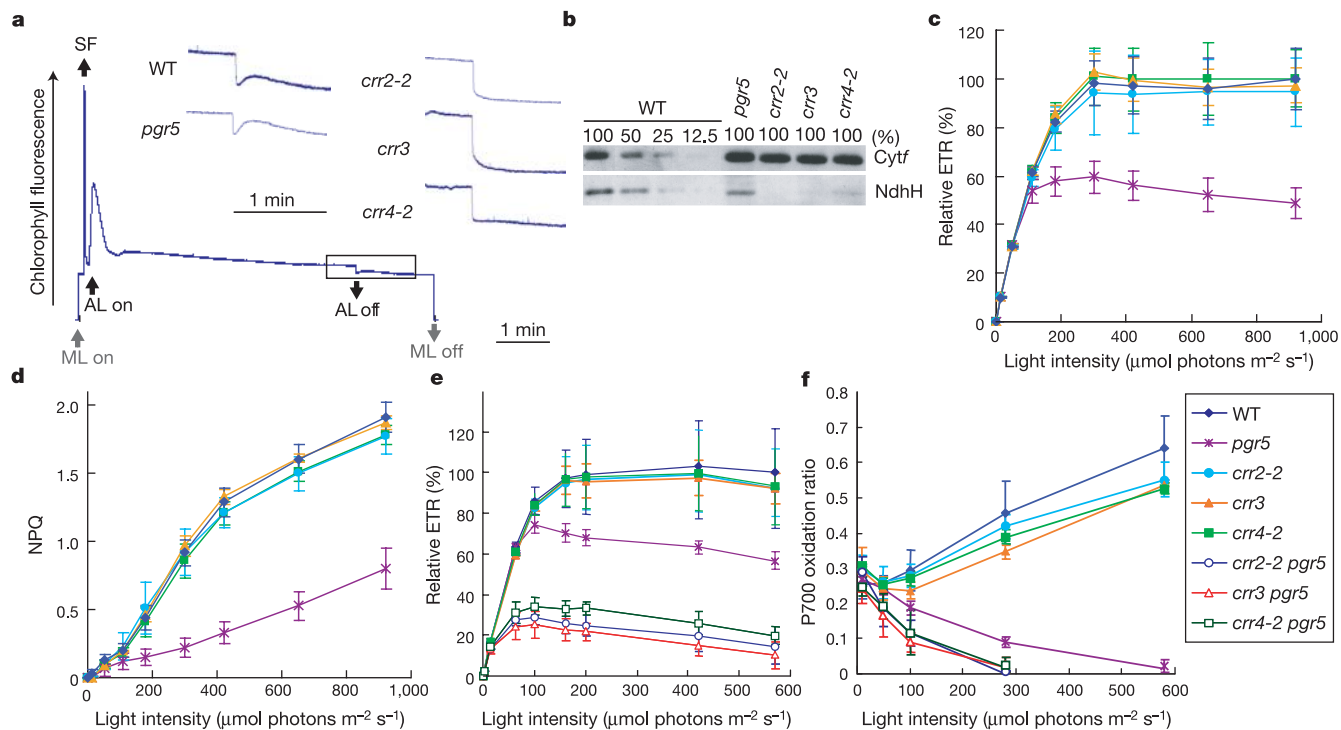


Figure 1 Characterization of single and double mutants. Seedlings were cultured at 50 $\mu\text{mol photons m}^{-2} \text{s}^{-1}$ (a–d) and 20 $\mu\text{mol photons m}^{-2} \text{s}^{-1}$ (e, f). **a**, Typical trace of chlorophyll fluorescence change in wild type. Vertical bars indicate the timing of on (upward) and off (downward) points of measuring light (ML), actinic light (AL) of 50 $\mu\text{mol photons m}^{-2} \text{s}^{-1}$, and a saturating light flash (SF, 3,000 $\mu\text{mol photons m}^{-2} \text{s}^{-1}$, 800 ms), respectively. Insets are magnified traces from the boxed area of transient increase in chlorophyll fluorescence due to NDH activity for the various genetic backgrounds. **b**, Immunodetection of the NDH (NdhH) and cytochrome *b₆f* (Cyt f) subunits, the latter of which was used as a loading control. The lanes were loaded with

0.2 μg chlorophyll for Cyt f and 2 μg chlorophyll for NdhH (100%), and a series of dilutions as indicated. **c**, Light-intensity dependence of the ETR. The ETR is depicted relative to a maximal value of $\Phi_{PSII} \times PFD$ (photon flux density, $\mu\text{mol photons m}^{-2} \text{s}^{-1}$) in the wild type of 100%. **d**, Light-intensity dependence of NPQ. **e**, Light-intensity dependence of the ETR. **f**, Light-intensity dependence of the oxidation level of P700 ($\Delta A/\Delta A_{max}$), where ΔA and ΔA_{max} are *in vivo* absorbance at 820 nm during actinic light and far-red light illumination, respectively. Each value represents mean $\pm$ standard deviation ($n = 5$) in c–f.

using chloroplast transformation impaired the chlororespiratory reduction of plastoquinone, which possibly functions in cyclic flow in the light^{14–16}. In contrast to the *pgr5* mutation, knockout of *ndh* genes did not affect overall photosynthetic electron transport; however, tobacco *ndh* knockout lines are sensitive to some environmental stresses, suggesting that the NDH-dependent cyclic flow is essential in photoprotection^{17,18}. The mechanisms through which the NDH-dependent pathway operates in photoprotection are yet to be determined.

What is the physiological function of redundancy in cyclic flow? To address this question we isolated *crr* (chlororespiratory reduction) mutants that have impaired NDH activity in *Arabidopsis*¹⁹. Because chloroplast transformation is not feasible in *Arabidopsis*, we isolated nuclear mutants affected in NDH activity by focusing on changes in chlorophyll fluorescence relating to NDH activity¹⁶. *crr* mutants were identified by their lack of a transient increase in chlorophyll fluorescence after actinic light illumination ceases (Fig. 1a). This change in fluorescence is due to the chlororespiratory reduction of plastoquinone by NDH activity and is absent in tobacco *ndh* knockout lines^{14–16}. The *crr2-2* mutant has a nonsense mutation in the gene encoding the PPR (pentatricopeptide repeat) protein, which functions in the intergenic processing of chloroplast RNA between *rps7* and *ndhB*. This processing is essential for the translation of *ndhB*, which encodes an NDH subunit¹⁹. To eliminate the possibility that *crr2-2* has pleiotropic defects that influence the phenotype of *crr2-2 pgr5*, we used two additional *crr* mutants, *crr3* and *crr4-2*. *crr4-2* is also defective in the function of a PPR protein involved in RNA editing and cannot create an initiation codon of the *ndhD* gene, which encodes another NDH subunit, whereas *crr3* contains a mutation affecting an unknown protein with predicted targeting to the thylakoid membrane (T.S. *et al.*, unpublished results). Reduced levels of the NDH complex (as estimated by NdhH levels) explain the lack of a transient increase in chlorophyll fluorescence after turning off actinic light (Fig. 1b). NdhH levels were reduced most markedly in *crr2-2* (to <6.25%) (ref. 19 and Fig. 1b) and to a lesser extent in *crr3* (approximately 12.5%) and *crr4-2* (12.5–25%).

The light-intensity dependence of the electron transport rate (ETR) is indicative of the relative amount of electrons passing through PSII during steady-state photosynthesis. This was not impaired at any light intensity in *crr* mutants (Fig. 1c), indicating that NDH-dependent cyclic flow is not essential in photosynthesis, thus corroborating previous findings in tobacco^{14,16}. However, the ETR was saturated at both a lower light intensity and a lower level in

pgr5 than in wild type, but it was unaffected at light intensities less than 100 $\mu\text{mol photons m}^{-2} \text{s}^{-1}$, which had previously been reported⁷.

Non-photochemical quenching (NPQ) is a chlorophyll fluorescence parameter that is roughly indicative of the level of thermal dissipation²⁰. Although the NDH complex functions in cyclic flow, NPQ induction was not affected in any *crr* mutants (Fig. 1d). This is consistent with observations for the tobacco *ndhB* knockout line¹⁶. In contrast, NPQ induction was severely affected in *pgr5* mutants, where it can be attributed to the insufficient generation of ΔpH at high light intensity.

To create double mutants lacking both activities of cyclic flow, *pgr5* was crossed with *crr2-2*, *crr3* and *crr4-2*, respectively. Although the chlorophyll content was slightly reduced in *pgr5* mutants, the growth rate was not affected in soil (Fig. 2; see also Supplementary Fig. 1). *crr* mutants did not display any visible phenotype; however, the growth rate and chlorophyll content were significantly reduced in the double mutants. These phenotypes were most evident in the double mutant *pgr5 crr2-2*, in which the accumulation of the NDH complex was most severely affected (Fig. 1b).

To assess the activity of cyclic flow, ruptured chloroplasts were isolated from double mutant, single mutant and wild-type plants. Cyclic flow activity was assayed by plastoquinone reduction upon addition of ferredoxin and NADPH to the chloroplast preparations. In this assay system, NADPH is essential for electron donation to ferredoxin through the reverse reaction of ferredoxin–NADP⁺ reductase²¹. Plastoquinone reduction was monitored as an increase in chlorophyll fluorescence emitted from PSII (Fig. 3a). Although NADPH did not reduce plastoquinone, it was rapidly reduced by subsequent addition of ferredoxin (Fig. 3b). The kinetics and the final reduced level were significantly lower in *pgr5* mutants. Addition of the ferredoxin–plastoquinone reductase activity inhibitor antimycin A mimicked the *pgr5* phenotype in wild-type plants but did not affect the *pgr5* phenotype further, as reported previously⁷.

In *crr2-2* seedlings, both the kinetics and the final level of plastoquinone reduction were affected to a similar extent as in *pgr5* seedlings. Addition of antimycin A to the *crr2-2* chloroplasts completely blocked the remaining activity of plastoquinone reduction. In the double mutant *crr2-2 pgr5*, plastoquinone reduction activity was completely impaired, indicating that the PGR5- and NDH-dependent activities account for almost all cyclic flow in *Arabidopsis*. As reported previously in tobacco²², NDH-dependent plastoquinone reduction activity requires ferredoxin as well as NADPH.

To characterize photosynthetic activity in the double mutants, the light-intensity dependence of ETR was analysed (Fig. 1e). Owing to the sensitivity of *crr2-2 pgr5* seedlings to the low light intensity of 50 $\mu\text{mol photons m}^{-2} \text{s}^{-1}$, all seedlings used in this assay were cultured at 20 $\mu\text{mol photons m}^{-2} \text{s}^{-1}$. The ETR was saturated at a lower level in wild-type seedlings cultured at 20 $\mu\text{mol photons m}^{-2} \text{s}^{-1}$ than in those cultured at 50 $\mu\text{mol photons m}^{-2} \text{s}^{-1}$ (Fig. 1c, e). The ETR was not affected at any light intensity in *crr*

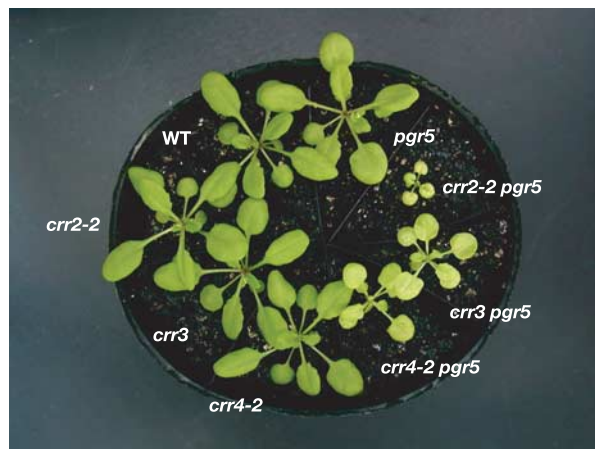


Figure 2 Visible phenotype of mutants. Seedlings were cultured at 50 $\mu\text{mol photons m}^{-2} \text{s}^{-1}$ for 5 weeks.

Table 1 Oxygen evolution in isolated ruptured chloroplasts

Genotype	Oxygen ($\mu\text{mol O}_2 \text{ mg}^{-1} \text{ chlorophyll h}^{-1}$)
Wild type	101 $\pm$ 7
<i>pgr5</i>	105 $\pm$ 11
<i>crr2-2</i>	93 $\pm$ 4
<i>crr3</i>	98 $\pm$ 6
<i>crr4-2</i>	92 $\pm$ 8
<i>crr3 pgr5</i>	91 $\pm$ 17
<i>crr4-2 pgr5</i>	101 $\pm$ 3

Photosynthetic oxygen evolution was measured with an electron acceptor (0.6 mM potassium ferricyanide). Values are mean $\pm$ s.d. ($n = 3$ –5).

mutants with seedlings cultured at either 50 or 20 $\mu\text{mol photons m}^{-2} \text{s}^{-1}$. The ETR was specifically affected at a high light intensity of more than 100 $\mu\text{mol photons m}^{-2} \text{s}^{-1}$ in *pgr5*; however, in the double mutants the ETR was severely affected even at the low light intensity of 50 $\mu\text{mol photons m}^{-2} \text{s}^{-1}$, and was saturated at a much lower level than in *pgr5*. This result correlates strongly with growth phenotypes of the double mutants at 50 $\mu\text{mol photons m}^{-2} \text{s}^{-1}$ (Fig. 2). We conclude that linear flow is severely affected even at very low light intensities in the double mutants. It is probable that the secondary effects of the reduced linear flow activity enhanced the phenotype through signals from the redox state of linear pathway²³ or excessive generation of active oxygen species²⁴.

Complete inhibition of cyclic flow severely affects linear flow *in vivo* (Fig. 1e). To assess the possible defect in the machinery of linear flow in the double mutants, electron transport to an artificial electron acceptor, ferricyanide, was evaluated by measuring O_2 evolution using isolated thylakoids (Table 1). We did not include the *crr2-2 pgr5* double mutant, which exhibits the most severe phenotype in this analysis, because PSII was photodamaged in these seedlings cultured even at 20 $\mu\text{mol photons m}^{-2} \text{s}^{-1}$ ($F_v/F_m = 0.696 \pm 0.02$ ($\pm \text{s.d.}$), where F_m = maximum fluorescence level in the dark, and $F_v = F_m - F_o$, where F_o is the minimum fluorescence level in the dark). This result suggests that the restriction of electron transport is partially irreversible in *crr2 pgr5* plants.

In spite of the severe reduction of electron flow activity *in vivo*, O_2 evolution dependent on linear flow by means of an artificial electron acceptor was not affected in any of the genetic backgrounds (Table 1). This result indicates that the machinery of linear flow was not directly impaired in the double mutants. We believe that the complete block of cyclic flow activity in the double mutants restricts linear flow in a reversible manner probably at the electron acceptors from PSI.

Although linear flow from water to ferricyanide was not directly impaired in ruptured chloroplasts (Table 1), it was severely affected *in vivo* (Fig. 1e). These results suggest that electron acceptance from PSI is limited in the double mutants. To characterize further the defect in electron transport, the oxidation level of the P700 reaction centre of PSI was monitored during steady-state photosynthesis (Fig. 1f). In the wild type, P700 oxidation increased at light intensities above 100 $\mu\text{mol photons m}^{-2} \text{s}^{-1}$ owing to the induction

of thermal dissipation and the restriction of electron transport at the cytochrome *b₆f* complex. In contrast, P700 oxidation declined as light intensity increased in *pgr5*, as previously reported⁷. The low P700 oxidation ratio in *pgr5* is caused by the reduction of P700^+ by electrons that have returned from a series of acceptor side electron carriers in PSI (A_0 , A_1 , F_X and $\text{F}_\text{A}/\text{F}_\text{B}$). This phenomenon is referred to as a charge recombination of P700 and is caused by over-reduction of electron acceptors from PSI (ferredoxin and NADP^+)^{7,17}. Although the ETR monitored at PSII was not affected, the P700^+ level was slightly lower in *crr* mutants than in wild type. Therefore, the P700^+ level is a more direct and sensitive indicator of the electron acceptance capacity from PSI. The deficiency in P700 oxidation was also evident in seedlings cultured at 50 $\mu\text{mol photons m}^{-2} \text{s}^{-1}$ (data not shown). These results indicate that the NDH complex also accepts electrons from PSI, although the efficiency is much less than that of ferredoxin-plastoquinone reductase. In the double mutants, P700 oxidation was lower than in *pgr5* even at a very low light intensity. These results indicate that the double defects in cyclic flow severely affect electron acceptance from PSI *in vivo*. Although this phenotype was observed in the single *pgr5* mutant at a high light intensity, it was evident at a very low light intensity in the double mutants.

Throughout a long history of interest, the physiological significance of cyclic flow has remained a matter of debate. We have previously shown that the PGR5-dependent pathway is essential to induce thermal dissipation under excessive light conditions and also to protect PSI from irreversible photodamage⁷. However, in the double mutants, the phenotype of severe reduction in linear flow activity even at a low light intensity (Fig. 1e) cannot be explained simply by a defect in photoprotection. The defect in the double mutants cannot be attributed to the reduced level of NPQ because electron transport is only marginally affected in the *Arabidopsis* mutant *npq4*, in which the induction of thermal dissipation is completely impaired²⁵. The linear flow activity was restored by adding ferricyanide to the ruptured chloroplasts (Table 1). This result indicates that electron acceptors from PSI (oxidized ferredoxin and NADP^+) are less available in the stroma of the double mutants, even at a low light intensity. This conclusion is supported by the lower levels of P700 oxidation *in vivo* (Fig. 1f). We have previously shown that P700^+ is restored to wild-type levels in *pgr5* by infiltrating leaf disks with an artificial electron acceptor from PSI (methyl viologen)⁷. We conclude that cyclic flow is essential for the prevention of stroma over-reduction.

How does cyclic flow protect the stroma from over-reduction? This question can be answered by applying the classical idea that cyclic flow contributes towards ATP synthesis during steady-state photosynthesis^{9–11}. During linear flow the reduction of NADP^+ and proton translocation across the thylakoid membrane is coupled, and hence, the ratio of ATP/NADPH generation is fixed (assuming that the Q-cycle in the cytochrome *b₆f* complex is functioning constantly). There has been much discussion about whether the ratio of ATP/NADPH generation satisfies the requirement of CO_2 fixation²⁶. We should take into account the fact that ATP is required in a variety of processes in the chloroplast in addition to CO_2 fixation. Alternative electron transport including cyclic flow and/or O_2 reduction by the Mehler reaction (water–water cycle)²⁷ may serve to adjust the ATP/NADPH ratio during photosynthesis^{28,29}. As evident from the NPQ in *pgr5*, cyclic flow contributes to ΔpH generation more significantly than we first thought. However, the severe phenotypes observed in double mutants indicate that we have still underestimated the physiological significance of cyclic flow. We believe that without the aid of cyclic flow, linear flow cannot maintain the correct ratio of ATP/NADPH production. This will ultimately lead to excessive accumulation of NADPH in the stroma and thereby, its over-reduction. □

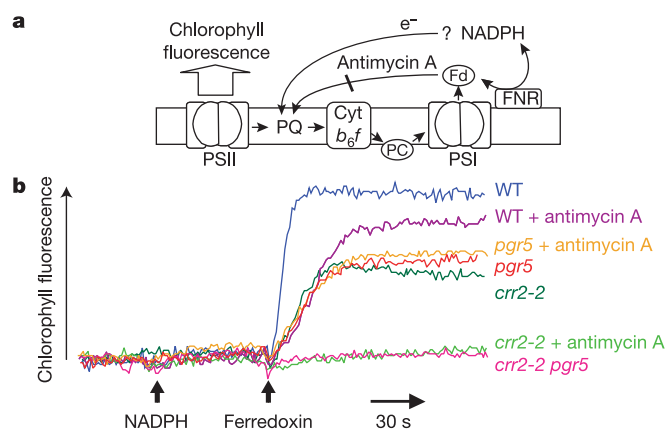


Figure 3 Electron donation to plastoquinone in ruptured chloroplasts. **a**, Schematic model of electron flow from NADPH to plastoquinone (PQ) via ferredoxin (Fd) in this assay. PC, plastocyanin; FNR, ferredoxin-NADP⁺ reductase. The electron donor to the NDH complex has not yet been determined (see text). **b**, Increases in chlorophyll fluorescence by addition of NADPH (0.25 mM) and ferredoxin (5 μM) under the illumination of weak measuring light (1.0 $\mu\text{mol photons m}^{-2} \text{s}^{-1}$) were monitored in osmotically ruptured chloroplasts (20 $\mu\text{g ml}^{-1}$) of the wild type (WT) and each mutant. Ruptured chloroplasts were incubated with 2 μM antimycin A, as indicated, before measurement.

Methods

Plant materials

Arabidopsis thaliana wild type (ecotype Columbia *g11*) and mutants were cultured in soil for 4–5 weeks under the following conditions: 50 $\mu\text{mol photons m}^{-2} \text{s}^{-1}$, 16/8 h light/dark cycle and 23 °C. Seedlings analysed in Figs 1e, f and 3 were cultured at a lower light intensity of 20 $\mu\text{mol photons m}^{-2} \text{s}^{-1}$.

Electron transport analysis

Chlorophyll fluorescence parameters and NDH activity were measured using a MINI-PAM (pulse-amplitude modulation) portable chlorophyll fluorometer (Walz), as previously described^{6,15,18}. Φ_{PSII} and NPQ were calculated as $(F_m' - F_s)/F_m'$ and $(F_m - F_m')/F_m'$, respectively, where F_m' is maximum fluorescence level in the light, and F_s is the steady-state fluorescence level. The redox change of P700 assessed by monitoring absorbance at 820 nm was measured with a PAM chlorophyll fluorometer (Walz) with an emitter-detector unit ED P700DW, as previously described^{6,30}.

Intact chloroplasts were isolated and osmotically ruptured as previously described⁷. Photosynthetic oxygen evolution was measured using a Clark-type oxygen electrode, with an electron acceptor (0.6 mM potassium ferricyanide) in the presence of 2 mM methyl amine at 2,200 $\mu\text{mol photons m}^{-2} \text{s}^{-1}$. Ferredoxin-dependent plastoquinone reduction activity was measured in ruptured chloroplasts²². We used 5 μM maize ferredoxin and 0.25 mM NADPH as electron donors.

Immunoblot analysis

Chloroplast protein was fractionated through 12% SDS–polyacrylamide gel electrophoresis and transferred to a polyvinylidene fluoride membrane. Immunodetection was performed with antibodies against NdhH and cytochrome *f* (Cyt*f*).

Received 21 January; accepted 19 April 2004; doi:10.1038/nature02598.

- Hill, R. & Bendall, E. Function of the two cytochrome components in chloroplasts—a working hypothesis. *Nature* **186**, 136–137 (1960).
- Heber, U. & Walker, D. Concerning a dual function of coupled cyclic electron transport in leaves. *Plant Physiol.* **100**, 1621–1626 (1992).
- Fork, D. C. & Herbert, S. K. Electron transport and photophosphorylation by photosystem I *in vivo* in plants and cyanobacteria. *Photosynth. Res.* **36**, 149–168 (1993).
- Bendall, D. S. & Manasse, R. S. Cyclic photophosphorylation and electron transport. *Biochim. Biophys. Acta* **1229**, 23–38 (1995).
- Joliot, P. & Joliot, A. Cyclic electron transfer in plant leaf. *Proc. Natl Acad. Sci. USA* **99**, 10209–10214 (2002).
- Müller, P., Li, X.-P. & Niyogi, K. K. Non-photochemical quenching: A response to excess light energy. *Plant Physiol.* **125**, 1558–1566 (2001).
- Munekage, Y. *et al.* PGR5 is involved in cyclic electron flow around photosystem I and is essential for photoprotection in *Arabidopsis*. *Cell* **110**, 361–371 (2002).
- Matsubayashi, T. *et al.* Six chloroplast genes (*ndhA-F*) homologous to human mitochondrial genes encoding components of the respiratory chain NADH dehydrogenase are actively expressed: Determination of the splice sites in *ndhA* and *ndhB* pre-mRNAs. *Mol. Gen. Genet.* **210**, 385–393 (1987).
- Arnon, D. I., Allen, M. B. & Whatley, F. R. Photosynthesis by isolated chloroplasts. *Nature* **174**, 394–396 (1954).
- Arnon, D. I., Tsujimoto, H. Y. & McSwain, B. D. Ferredoxin and photosynthetic phosphorylation. *Nature* **214**, 562–566 (1967).
- Tagawa, K., Tsujimoto, H. Y. & Arnon, D. Role of chloroplast ferredoxin in the energy conversion process of photosynthesis. *Proc. Natl Acad. Sci. USA* **49**, 567–572 (1963).
- Ogawa, T. A gene homologous to the subunit-2 gene of NADH dehydrogenase is essential to inorganic

- carbon transport of *Synechocystis* PCC6803. *Proc. Natl Acad. Sci. USA* **88**, 4275–4279 (1991).
- Mi, H., Endo, T., Ogawa, T. & Asada, K. Thylakoid membrane-bound pyridine nucleotide dehydrogenase complex mediates cyclic electron transport in the cyanobacteria *Synechocystis* PCC 6803. *Plant Cell Physiol.* **36**, 661–668 (1995).
- Burrows, P. A., Sazanov, L. A., Svab, Z., Maliga, P. & Nixon, P. J. Identification of a functional respiratory complex in chloroplasts through analysis of tobacco mutants containing disrupted plastid *ndh* genes. *EMBO J.* **17**, 868–876 (1998).
- Kofer, W., Koop, H.-U., Wanner, G. & Steinmüller, K. Mutagenesis of the genes encoding subunits A, C, H, I, J and K of the plastid NAD(P)H-plastoquinone-oxidoreductase in tobacco by polyethylene glycol-mediated plastome transformation. *Mol. Gen. Genet.* **258**, 166–173 (1998).
- Shikanai, T. *et al.* Directed disruption of the tobacco *ndhB* gene impairs cyclic electron flow around photosystem I. *Proc. Natl Acad. Sci. USA* **95**, 9705–9709 (1998).
- Endo, T., Shikanai, T., Takabayashi, A., Asada, K. & Sato, F. The role of chloroplastic NAD(P)H dehydrogenase in photoprotection. *FEBS Lett.* **457**, 5–8 (1999).
- Horváth, E. M. *et al.* Targeted inactivation of the plastid *ndhB* gene in tobacco results in an enhanced sensitivity of photosynthesis to moderate stomatal closure. *Plant Physiol.* **123**, 1337–1350 (2000).
- Hashimoto, M., Endo, T., Peltier, G., Tasaka, M. & Shikanai, T. A nucleus-encoded factor, CRR2, is essential for the expression of chloroplast *ndhB* in *Arabidopsis*. *Plant J.* **36**, 541–549 (2003).
- Krause, G. H. & Weis, E. Chlorophyll fluorescence and photosynthesis: The basics. *Annu. Rev. Plant Physiol. Plant Mol. Biol.* **42**, 313–349 (1991).
- Miyake, C. & Asada, K. Ferredoxin-dependent photoreduction of the monodehydroascorbate radical in spinach thylakoids. *Plant Cell Physiol.* **35**, 539–549 (1994).
- Endo, T., Shikanai, T., Sato, F. & Asada, K. NAD(P)H dehydrogenase-dependent, antimycin A-sensitive electron donation to plastoquinone in tobacco chloroplasts. *Plant Cell Physiol.* **39**, 1226–1231 (1998).
- Pfannschmidt, T., Nilsson, A. & Allen, J. F. Photosynthetic control of chloroplast gene expression. *Nature* **397**, 625–628 (1999).
- op den Camp, R. G. L. *et al.* Rapid induction of distinct stress responses after the release of singlet oxygen in *Arabidopsis*. *Plant Cell* **15**, 2320–2332 (1999).
- Li, X.-P. *et al.* A pigment-binding protein essential for regulation of photosynthetic light harvesting. *Nature* **403**, 391–395 (2000).
- Allen, J. Photosynthesis of ATP-electrons, proton pumps, rotors, and poise. *Cell* **110**, 273–276 (2002).
- Asada, K. The water–water cycle in chloroplasts: Scavenging of active oxygens and dissipation of excess photons. *Annu. Rev. Plant Physiol. Plant Mol. Biol.* **50**, 601–639 (1999).
- Backhausen, J. E., Kitzmann, C., Horton, P. & Scheibe, R. Electron acceptors in isolated intact spinach chloroplasts act hierarchically to prevent overreduction and competition for electrons. *Photosynth. Res.* **64**, 1–13 (2000).
- Heber, U. Irrungen, Wurrungen? The Mehler reaction in relation to cyclic electron transport in C3 plants. *Photosynth. Res.* **73**, 223–231 (2002).
- Schreiber, U., Klughammer, C. & Neubauer, C. Measuring P700 absorbance changes around 830 nm with a new type of pulse modulation system. *Z. Naturforsch.* **43C**, 686–698 (1988).

Supplementary Information accompanies the paper on www.nature.com/nature.

Acknowledgements We thank M. Miyata and E. Habe for their technical assistance, and T. Mae, A. Makino and G. Peltier for antibodies. A. Portis is acknowledged for his critical reading of the manuscript. T.S. was supported by grants from JSPS.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to T.S. (shikanai@agr.kyushu-u.ac.jp).

Contacts

Publisher: Ben Crowe

Editor: Paul Smaglik

Marketing Manager: David Bowen

European Head Office, London

The Macmillan Building
4 Crinan Street
London N1 9XW, UK
Tel +44 (0) 20 7843 4961
Fax +44 (0) 20 7843 4996
e-mail: naturejobs@nature.com

Naturejobs Sales Director:

Nevin Bayoumi (4978)

UK/ RoW/ Ireland:

Matt Powell (4953)

Andy Douglas (4975)

Frank Phelan (4944)

Scandinavia/ Spain/ Portugal:

Evelina Rubio Håkansson (4973)

Natureevents: Silke Opstrup (4994)

France/ Switzerland:

Amelie Pequignot (4974)

Production Manager:

Billie Franklin
To send materials use London
address above.

Tel +44 (0) 20 7843 4814

Fax +44 (0) 20 7843 4996

e-mail: naturejobs@nature.com

Naturejobs web development:

Tom Hancock

Naturejobs online production:

Stefan Hales

European Satellite Office

Germany/ Austria/ Italy/

The Netherlands/ Belgium:

Patrick Phelan, Odo Wulffen

Tel + 49 89 54 90 57 11/-2

Fax + 49 89 54 90 57 20

e-mail: p.phelan@nature.com

o.wulffen@nature.com

US Head Office, New York

345 Park Avenue South,
10th Floor, New York, NY 10010-1707
Tel +1 800 989 7718
Fax +1 800 989 7103
e-mail: naturejobs@natureny.com

US Sales Manager:

Peter Bless

Japan Head Office, Tokyo

MG Ichigaya Building (5F),
19-1 Haraikatamachi,
Shinjuku-ku,
Tokyo 162-0841
Tel +81 3 3267 8751
Fax +81 3 3267 8746

Asia-Pacific Sales Director:

Rinoko Asami
e-mail: r.asami@naturejpn.com

naturejobs

Never-ending stories

At the beginning of this year, *Naturejobs* introduced you to four new writers — graduate students who penned a monthly journal to provide you with an insight into the highs and lows of building a career in science. Six months on, their stories have undergone many twists and turns, and our correspondents have wrestled with dilemmas that will be familiar to many.

As he worked to complete his thesis, AIDS researcher Tshaka Cunningham at Rockefeller University in New York was weighing up the merits of doing his postdoc abroad. Unusually for a US-born scientist he chose to leave the Big Apple and head for France. Amber Jenkins, meanwhile, found herself flitting between Imperial College London and Fermilab's particle accelerator in Batavia, Illinois, as she chased data for her PhD. She has used these extended trips to evaluate whether her career should take her back home to Britain when she finishes her thesis.

Materials scientist Sidney Omelon at Mount Sinai Hospital in Toronto, Canada, is still wondering whether she should return to industry for a third time, stay with academia or pursue an alternative career. And Philipp Angerer at the Swiss Federal Institute of Technology in Zurich is trying to decide whether to join a company or start one after he defends his thesis.

A common theme for all four writers has been the need for balance between the lab and the outside world. They have also sometimes had to strike a balance between being mentors to younger co-workers and getting advice from their senior colleagues. Each instalment of these stories can now be found at the *Naturejobs* Graduate Channel (www.nature.com/naturejobs/channels/graduate/graduate-journal.html), so readers who missed early episodes can tune in. Whether you are just starting your scientific career, or nearing the end, these stories are likely to have you nodding your head in recognition.

Paul Smaglik
Naturejobs editor



Contents

CAREER VIEW

Nuts & Bolts

Introductory matters

Graduate Journal

The final straight

Movers

Andre Pernet

p584

WWW.NATUREJOBS.COM

Career centre

Information on the
scientific job market

FOCUS

SPOTLIGHT

RECRUITMENT

ANNOUNCEMENTS

EVENTS

GRADUATE JOURNAL

The final straight

I did a crazy thing this week. I signed up to run a half-marathon in four months' time. Me, a complete novice. Someone who hates running. Why? Because, crazy as I am, I love a challenge. They say running is all about mental discipline. It's just as well, because it's not the only race I'll be running in the next few months. I'm also entering the final straight of my PhD.

At this stage, the end is in sight. But getting there is still going to be a slog. The view is rather daunting, but I know I have to keep a cool head and take one step at a time.

In truth it's an exciting period in my PhD, a time of consolidation. Things are supposed to come together as a cohesive whole, although it sometimes feels like a frenzied rush. Will the experiment deliver enough data by the time I have to submit? I don't know, but that's the unpredictability of research. It doesn't always fit to a neat timetable and I may have to adapt, depending on what happens. At the same time I'm faced with major career decisions.

I think you have to trust that it will be all right. Writing up is a test of many things, chiefly patience, wits and sanity. I'm in for the long haul. It's too late to turn back now, but I believe it will be worth it when I cross that finishing line. ■

Amber Jenkins is a graduate student in particle physics at Imperial College, London, doing thesis research at Fermilab in Batavia, Illinois.

NUTS & BOLTS

Introductory matters

You only have one chance, as the saying goes, to make a good first impression. Because their work is so complex, most scientists could use a little help in simplifying their message when introducing themselves and describing their research.

Scientists often find themselves called on to work with peers in other disciplines, as well as with non-scientists in business or administration, and with the general public. To the uninitiated, detailed descriptions of your work may be incomprehensible.

Avoid the 'blah, blah, blah' syndrome by working on an introduction that is concise and memorable. The ability to highlight yourself and your work in an introduction can lead to job offers, funding and generally more engaged conversations.

Starting with your name, make sure your introduction



With Deb Koen
Careers consultant

is heard and understood. If it's a long or unusual name, you may have to offer a clue to help the listener remember it. For example, "Karin Smith: that's pronounced Karin, like 'car in the garage'." Repetition also helps, as in "Good afternoon. My name is Joseph. Joseph Lonnergon."

Once past the name, you are on to what you do: a brief, understandable and memorable description. Here, think 'talent show'. Focus on a skill (talent) and an example of it (show). Mention something you have just completed or are involved in now, such

as a project or study.

Take an example from last week's *Naturejobs* (see *Nature* 429, 484–485; 2004). Virologist John Alderete might introduce himself in the following way: "Hi, I'm John. John Alderete. I'm co-founder of a biotech start-up that recently created a test for a sexually transmitted disease and took it through to approval by the Food and Drug Administration." Repeating the first name will help the listener remember, and the 'talent show' statement is concise yet meaningful.

Although this approach may initially strike you as contrived, you will be sold once you experience the resulting improvement in communication. A clear and concise introduction of yourself goes a long way in opening the door to more meaningful conversation. ■

Deb Koen is vice-president of Career Development Services and a columnist for *The Wall Street Journal's CareerJournal.com*.

MOVERS Andre Pernet, president, Quark Biotech, Fremont, California



Andre Pernet had a dual education in science and business. The biochemist, who last year took over as president of Quark Biotech in Fremont, California, earned his MBA and PhD more or less at the same time. Since then he has put his skills to good use, carving a career path that has taken him from the lab bench to the corporate boardroom.

Most of this transition took place at Abbott Laboratories in Montreal, Canada, and Chicago, Illinois, where he worked for 26 years. But his abilities —

and his mettle — were tested when he became chief executive of Genset, a French biotech firm near Paris.

When he arrived in 2000, he found a company that had several divisions doing interesting science. But the goals of many of these sections bore scant resemblance to the core technology of the company of gene and protein identification. As Genset had yet to bring a product to market, the company was losing money.

To turn the situation around, Pernet sought buyers for those interesting but unprofitable components. "In the end we had a company that was coherent and had a lower cash-burn rate," he says.

But that experience was double-edged. "Being able to transform the company into something that is viable and successful is the most exciting part of the job," Pernet says. But it also meant laying off people, a process that

Pernet describes as "not so fun — but you've got to do it".

Pernet says that being involved in different phases of drug development has taught him how the whole business works. In his early years at Abbott, his main goal was to create chemicals with different but specific properties. As he rose through the ranks there, he learned how to manage the increased complexity of trials that involve human subjects.

And first at Genset and now at Quark, he says that communication skills are most important — how to get diverse teams of people to talk about basic and clinical problems with the aim of achieving specific business goals.

Quark is currently making a loss, but Pernet hopes that, with good communication skills, he can help turn it into a profitable company. The business lessons he learned at Genset should make that transformation easier for him. ■

CV

2000–02: Chief executive, Genset, Evry, France

1973–99: Abbott Laboratories, Montreal, Canada, and Chicago, Illinois (joined as a research chemist, rising to divisional vice-president, worldwide development of pharmaceuticals in 1992)

1973: MBA, McGill University, Montreal, Canada

1973: PhD in chemistry, University of Montreal, Canada